Citation for published version (APA):
Contents

Why Do Some Research Articles Receive More Online Attention? Reasons for Online Success as Measured with Altmetrics .......................................................... 1

Can Twitter Increase the Visibility of Chinese Publications? ............................................ 10

Exploratory Analysis on the Construction of Journal Evaluation Model Using Altmetrics Indices ........................................................................................................... 20

Alternative Metrics Correlations: Do Academic Reviews Correlate with Library Holdings? ................................................................................................................. 31

An Analysis of Scientific Co-author Network of Virtual Technology in China .................. 44

A Weighted Method for Citation Network Community Detection ...................................... 58

Research on Domain Knowledge Network Based on Bibliometrics ....................................... 68

Detecting Social Communities in Spanish Theses Defense Committee in the Area of Computer Science ........................................................................................................ 83

Publication Patterns in the Social Sciences and Humanities in Flanders and Poland ........ 95

Assessment Criteria for Early Career Researcher’s Proposals in the Humanities ............ 105

Cognitive and Organizational Classification of Publications in the Social Sciences and Humanities .............................................................................................................. 112

Peer Review as A Delineation Criterion in Data Sources for the Assessment and Measurement of Scholarly Book Publishing ......................................................... 118

Bibliometric Analysis of Publications from Post-Soviet Countries in Psychological Journals in 1992–2016 ........................................................................................................ 125

Global Overview of Unmanned Aerial Vehicles Research: Country-level and Organisation-level Bibliometric Analysis ........................................................................... 136

A Parameter-free Index for Identifying Under-cited Sleeping Beauties in Science ........... 148
How is Citespace Used and Cited in the Literature? An Analysis of the Articles Published in English and Chinese Core Journals ................................................................. 158

China as Number 1? A Case for China Regaining World Leadership of Science and Technology .................................................................................................................. 166

A Peer Review Method Based on Identify Experts’ Guanxi .............................................. 178

The small-world phenomenon of microblog communication networks in China: an empirical study based on complex network analysis ......................................................... 186

A bibliometric analysis of articles published by Chinese authors in library and information science journals of SSCI .................................................................................................... 194

Impact and usage indicators for the assessment of research in scientific disciplines and journals .......................................................................................................................... 203

A New Approach to Evaluate the Impact of Papers with Combining Social and Academic Influence .................................................................................................................. 214

Scholars on Twitter: Who and How Many Are They? .................................................... 224

The Usability of Altmetrics in Academic Evaluation ......................................................... 236

Identifying and tracking scientific and technological knowledge memes from citation networks of publications and patents ................................................................. 248

Event Detection in Scientific Mapping based on a Novel Structural Community Similarity Algorithm .................................................................................................................. 258

Networks of international collaboration and mobility: a comparative study ....................... 270

Study on Ubiquitous Network Intelligent Service Based on KDD ...................................... 281

How large is large enough? ................................................................................................ 288

Do Mathematicians, Economists and Biomedical Scientists Trace Large Topics More Strongly Than Physicists? .................................................................................................. 300
Using Machine Learning to Identify Novel Awards-NSF Material Awards as A Case Study ................................................................. 312

Measuring Scientific Knowledge Flows by Deploying Citation Context Analysis using Machine Learning Approach on PLoS ONE Full Text .......................................................................................... 322

Effect of publication month on citation impact ...................................................................................................................... 334

Do under-cited influential sleeping beauties exist? ..................................................................................................................... 344

Aerospace Discipline Study Based on Highly Cited Papers ........................................................................................................ 350

Differences in citation personal display: Does it exist in three social science disciplines? ....................................................... 360

Usage Pattern Analysis of Academic Articles from Two Chinese Journals .................................................................................. 366

Research Progress of Library and Information Science: Exploring Chinese Scientific Articles from 2001 to 2015 ......................................................................................................................... 376

Disciplinary Differences in the Achievements of the National Natural Science Award in China ................................................................................................................................. 388

Evaluating Effects of World-Class University Program Using Propensity Score Matching Evidence from China’s Project 211 ........................................................................................................ 400

How visible is the research of different countries? An analysis of global vs. local reach of WoS publications on Twitter ...................................................................................................................... 412

The diffusion of medical knowledge in social media - An empirical investigation based on PLOS ALM data ............................................................... 420

An Integrated Analysis of Topical and Emotional Evolution of Microblog Public Opinions on Public Emergencies .............................................................................................................. 431

Communicating Scientific Video Articles on Twitter: An Initial Exploration of JoVE Publications ......................................................................................... 442

Normalization of zero-inflated data: An empirical analysis of a new indicator family ................................................................. 448

M-Score: An Indicator Quantifying Individual's Scientific Research Output ...................................................................................... 460
Discrimination Measurement Method on H-index and G-index Using Jain’s Fairness Index .......................................................... 466

Allocating the Credit in High Energy Physics Collaborative Research .......................... 477

Do Patent Citations to Conference Papers Differ from Journal Articles? Evidence of Technological Impact through Google Patents ...................................................... 490

Impact of multidisciplinary research on innovation ....................................................... 502

Measuring and Forecasting Technology Market Potential ........................................... 514

Detecting Science & Technology of Lockheed Martin Corporation throughout Funded Papers .............................................................................................................. 525

Exploring the Similarity of Articles Co-cited at Different Levels ............................... 549

Is the Relationship between the Impact Factor and Papers’ Citations Really Weakening? .................................................................................................................. 561

UK ethnic minority cancer researchers: their origins, destinations and sex ............... 568

Multiple-keyword Co-occurrence Analysis and Time Series Prediction of Research Hotspots .............................................................................................................. 580

European newspaper reports of non-communicable disease research, 2002-13 ....... 587

Main Path Identification involving Article’s Associated Attribute: A Case Study of Synthetic Biology ........................................................................................................... 599

The effect of “open access” on journal impact factors: A causal analysis of medical journals .............................................................................................................. 611

Analysis of highly cited papers in Rheumatology ......................................................... 622

Retraction: The “Other Face” of Research Collaboration ............................................. 632

Temporal Characteristics of Retracted Articles – Research in Progress ................. 650

Measuring Model of Collaboration Ability: the Collaborative Rate, Collaborative Breadth and Collaborative Depth ................................................................. 656

Some Reflections to China’s International Collaboration ........................................... 668
Multi-Scale H-Index: A New Measure to Assess the Scientific Impact of Scholars 680

A principled methodology for comparing relatedness measures for clustering publications 691

A new diversity indicator based on a similarity treatment expansion of the Gini coefficient 703

Research preferences of the G20 countries: a bibliometrics and visualization analysis 709

Bibliometric Study of Interdisciplinary Relations of Converging Technologies (Nano-Bio-Info-Cogno) 721

Mapping the development trajectory of 3D printing technologies 732

Knowledge Spillovers among Semiconductor Companies with Different Technology Positions and with Different Roles in the Industry Chain 741

The Missing Link Effect: A Case Study Using Patent Main Path Analysis 753

Co-citation analysis based on the position: a case study of Journal of Informetrics 766

Sleeping Beauty awakened by self-citation of a review: a case study of Judah Folkman hypothesis on angiogenesis 778

Reference behavior in the full text of scientific articles: A large-scale analysis 787


Evaluation of Scholarly Impact from an Integrated Perspective of Three-Dimensional Citations: A Case Study of Gene Editing 810

Medical Research Breakthroughs and the Public Engagement 817

Neologisms and affiliation name: a bibliometric perspective on the institutionalization of interventional radiology 827

Scientometrics for informing priority setting in health research: the case of mental and behavioural disorders 833

Cross-boundary collaboration of highly-cited scientists: A bibliometric study on scientific publication in agricultural science 840
A scientometric method for assessing an institution’s scientific collaboration policy 856

Revealing existing and potential partnerships: affinities and asymmetries in international collaboration and mobility 869

The structure and evolution of scientific collaboration from the perspective of symbiosis 881

A Brief Analysis on Top Scientists in the Field of Economics & Business: Based on Essential Science Indicators Database 896

Measuring and analysing the internal, topical coherence of Web of Science Subject Categories 902

Comprehensiveness and Overlap in Open Access Systems: Search Engines, Aggregate Institutional Repositories and Physic Open Sources 908

An Author Co-citation Analysis Method Combined with Metadata in Full Text 916

Research on Multiple Technology Paths Extraction Method Based on Semantics: Take Magnetic Head of Hard Disk Drive as Example 928

Relationships between Venture Capital and Listed Firms’ Technology Innovation: in the View of Patent, Growth Stage, Industries and Investor’s Nature 938

The choice of examiner citations for refusals: Evidence from the trilateral offices 950

Interactive overlay maps and technological field evolutions for US patent (USPTO) data 958

Productivity versus citation impact: A study of persons, not just authors 970

On the citation gap of articles naming countries 976

Identifying Funding Allocation Gaps Within an Academic Domain: Case Study of Robotics Research 982

A Logical Comparison of Citation Wakes and Generations 989

A novel analysis on global environmental quality research in the last two decades based on
bibliometric modeling indicators 995

Overlapping Thematic Structures Extraction with Mixed-Membership Stochastic Blockmodel 1007

A bootstrap analysis for finite populations 1013

Further examples of dominance structure measurement 1024

An analysis of international research collaboration based on research project data 1035

The Evolution of China’s Role in the International Scientific Collaboration Network 1052

Interdisciplinarity and collaboration: On the relationship between disciplinary diversity in the references and in the departmental affiliations 1064

Russian Index of Science Citation: Overview and Review 1076

Accuracy of citation data in Web of Science and Scopus 1087

Faculties Activity Research based on Local and Global Databases. Case Study 1093

Evaluating efficiency of digital databases used for scientific production in Chinese universities 1095

Funding Map for Research Project Relationships using Paragraph Vectors 1107

LitStoryTeller: an Interactive System for Visual Exploration of Scientific Papers Leveraging Named entities and Comparative Sentences 1118

University rankings in computer science: a study and visualization of ‘geo-based’ impact and conference proceeding (CORE) scores 1131

Using full-text data to create improved term maps 1136

Gender differences in research diversification behavior 1142

Do Google Scholar and Web of Science reflect women’s and men’s scholarly impact differently? A comparison of U.S. researchers in sociology and economics 1156
On a trajectory towards parity: an historical analysis of gender in funding from the National Science Foundation

Women in the Shadow of Big Men: The Case of Canada Excellence Research Chairs

An extension of the Characteristics Scales and Scores: isolating ‘inter-institutional’ from ‘intra-institutional’ variance in highly skewed real population distributions

The granularity of disciplinary structures for benchmarking citation impact. The case of CSS profiles

What factors are associated with more authors? A case study of Danish Economics

Statistical methods for forecasting “.it” domain names

On the Applicability of Altmetrics in the Evaluation of Scientific Journals

Does Monetary Support Increase the Number of Scientific Papers? An Interrupted Time Series Analysis

Topic based Research Competitiveness Evaluation

Open access papers: their growth over time and from different countries, and their citations

The Next Generation (Plus One): An Analysis of Doctoral Students’ Academic Fertility Using a Novel Approach for Identifying Advisors

Are Great Researchers Terrible Teachers? How research and teaching performance relate at U.S. universities

Research or management? An investigation of the impact of administrative roles on the research performance of academic administrators

Authorship, inventorship and division of labor in innovative research: an analysis of paper-patent pairs

Position Library and Information Science in the Field of Reading Research
Path to Success: An analysis of US educated elite academics in the United States 1314

Research on potential knowledge structure of international scientometrics 1326

Mapping the Semantic Word Shifts in Topics in the Field of Information Retrieval 1335

Word semantic change: The law of differentiation vs. the law of parallel change 1342

CS-LAS: A Scientific Literature Retrieval and Analysis System Based on Term Function Recognition (TFR) 1346

Incidental or influential? – A decade of using text-mining for citation function classification 1357

A Comparative Investigation of Citation Content Characteristics between Academic Monographs and Papers 1368

Modeling Study of Knowledge Diffusion in Scientific Collaboration Networks based on Differential Dynamics 1379

Research of cross-discipline based on Web-of-Science Categories 1394

Consistency of interdisciplinarity indicators? 1406

Interdisciplinary Knowledge Flow: across Disciplines and Scientists in Scientific Funding 1418

The association between university research indicators and success rates in the European Framework Programmes 1430

From academia to citizenry. Study of the flow of scientific information from R&D projects to scientific journals and social networks 1442

Assessing the Interdependencies between Scientific Disciplinary Profiles at the Country Level: a Pseudo-Likelihood Approach 1448

Multidimensional assessment of R&D performance: evidence from the pilot evaluation exercise of Russian public research institutions 1460
Measuring knowledge exchange of internationally mobile scientists - A bibliometric approach based on similarity

Journals as Communication Channels between National Sub-Communities

Author's Publication Strategies and Citation Distributions in Journals

Scientometric Research of Knowledge Communication on Social Media - A Case Study of Biomedical Science on Baidu Baike

Are bibliometrics, scientometrics, and informetrics different?

Webometrics of Think Tanks in Sub-Saharan Africa

Bibliometrics to Knowledgometrics: Theory Method and Principle of Metrics Science

Assessing Research and its Impacts: The Generalized Implementation Problem and a Doubly-Conditional Performance Evaluation Model

Methodological Challenges for the Comparison of Results of Topic Extraction from Scientific Literature

Study on the Method of Detecting the Research Fronts in the View of Topic Model

Integrating domain ontologies with topic modelling for mapping societal problems

Bibliometric study on big data research: An integration of topic model and citation network analysis

New Citation Analysis Perspective Based on Ontology and Linked Data

An Evaluation Model for Assessing Open Government Data in China

BIBLIOMETRIC ANALYSIS AND RANKING OF LIBRARY AND INFORMATION SCIENCE (LIS) RESEARCH AND PUBLICATIONS IN AFRICA

A Rainbow at the Skyline after the Storm of Indicators for Ranking Scientists
An Evolutionary Analysis of Research Integrity Based on Multidimensional Informetrics

Do under-cited influential sleeping beauties exist?

Which Drives Which? The Causal Relationship between Number of Editorial Board Members and Scientific Output of Universities in the Chemistry Field: a Granger Causality Test

Ternary Co-occurrence Latent Semantic Vector Space Model

Hybrid Relation Analysis (HRA): Methods and a Case Study, Brain Cancer


Prediction Model of Future Academic Impact for Young Scientists

Is the Gap in Scientific and Technological Strength Between G7 and BRICS Becoming Smaller or Larger?

To Be or Not to Be: Will Scientific Writing Affect Scientific Impact?

A new method to quantify the spatiotemporal dynamic of the academic papers using centroid method

Faculties Activity Research based on Local and Global Databases Case Study

Always Gold Glitters: When Accepted Papers Meet Processing Delay

MOOC Content Design: Take “Webometrics & Scientific Evaluation” Course as an Example

Statistical methods for forecasting “.it” domain names

Are the social sciences from the European post-socialist countries integrated in the "Western social sciences"?

An Artificial Neural Network Model Based on Altmetrics Indicators and Citation Counts
Using Full-text to Evaluate Impact of Different Software Groups ........................... 1666

Specialization and excellence in Latin American and Caribbean science: assessment proposal .................................................................................................................. 1668

On normalization by division and comparing citation counts .................................. 1670

Predicting the Future Impact of Scholars through Time-aware Academic Networks ··· 1672

Using Readership to Measure Interdisciplinarity ......................................................... 1674

Trial for the Data-oriented management using data-set of Equipment sharing in Hokkaido University ........................................................................................................ 1676

Computing the Influence of Disciplinary Keywords Based on h-Index ....................... 1678

Correlation between scientific production, energy production and funding on renewable energies at the global level .................................................................................. 1680

Publication in arXiv: a current bibliometric analysis of preprint publication benefits and drawbacks in science and technology ................................................................. 1682

"How bibliometrics is related with other information science topics. An approximation from a review journal!" ............................................................................................... 1684

Green Patent Value Evaluation Based on Social Network Analysis ............................ 1686

A scientometric method for assessing an institution’ scientific collaboration policy ·· 1697

An Altmetrics Study of TOP100 Samples in 2016 ........................................................ 1710

Analyzing the efficiency of research grants: a case study of grants issued by the National Natural Science Foundation of China ........................................................................ 1719

The evolution of business model research (1993-2016): based on a coword and maximum spanning tree analysis ......................................................................................... 1721

Research of the Role of Leading Enterprises in Promoting Industrial Technology Development
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>based on Patent Content Analysis</td>
<td>1737</td>
</tr>
<tr>
<td>Network Community-based Technological Cooperation Identification</td>
<td>1754</td>
</tr>
<tr>
<td>Overcoming the difficulties of enhancing and interpreting results of underlying topic visualization in texts</td>
<td>1766</td>
</tr>
<tr>
<td>Topic Detection and Evolution Analysis of Research Project based on LDA—A case Study of Projects on Ocean Acidification</td>
<td>1772</td>
</tr>
<tr>
<td>Bundle of sleeping-beauties: The case of Paul Hagenmüller and solid-state chemistry</td>
<td>1786</td>
</tr>
<tr>
<td>Web of Science™ as a Research Dataset</td>
<td>1788</td>
</tr>
<tr>
<td>Developing Armenian Science Citation Index: Obstacles and Challenges</td>
<td>1791</td>
</tr>
<tr>
<td>Analyzing the citation impact of scholarly books based on BKCI</td>
<td>1793</td>
</tr>
<tr>
<td>Topic modelling based network maps in cardiovascular research</td>
<td>1795</td>
</tr>
<tr>
<td>Science and Technology Indicators of Microalgae-Based Biofuel Research</td>
<td>1797</td>
</tr>
<tr>
<td>The Dissemination of the Concept of “Interventional Radiology” in Medical and Scientific Literature</td>
<td>1799</td>
</tr>
<tr>
<td>The Influence of Advance Online Publication on Journal Impact Factor</td>
<td>1801</td>
</tr>
<tr>
<td>A scaling approach to tackle the heterogeneity of HEIs</td>
<td>1805</td>
</tr>
<tr>
<td>Bibliometric Research of Faculties based on Local Databases</td>
<td>1807</td>
</tr>
<tr>
<td>Some Reflections to China’s International Collaboration</td>
<td>1809</td>
</tr>
</tbody>
</table>

BIBLIOMETRIC ANALYSIS AND RANKING OF LIBRARY AND INFORMATION SCIENCE (LIS) RESEARCH AND PUBLICATIONS IN AFRICA | 1821 |
Ternary Co-occurrence Latent Semantic Vector Space Model 1825

Scientometric Study on Technology Promoting Science Development 1827

Analysis on the Developments of Archival Appraisal Theory Based on Scientometrics (1990-2016) 1841

Classifying Patents by Tracing the Chronology of Patent Citation Increments 1855

Is the Gap in Scientific and Technological Strength Between G7 and BRICS Becoming Smaller or Larger? 1864

Nanotechnology Research Innovation and Commercialisation: Some Salient Aspects 1866

Patent Citation Inflation: The Phenomenon and its Measurement Methods 1878

Using Digital Traces as Metrics to Understand Value in the Scholarly Process 1890

Identifying and Visualizing Scientific Dynamics at Multiple Levels of Granularity 1896

On normalization by division and comparing citation counts 1901

Indicators for the Evaluation of Technological Activity 1903

Applying the method of reflections to scientific production in the EU 1905

How many keywords do authors assign to research articles – a multidisciplinary analysis? 1907

Investment and results in R&D in the temporal context of the economic crisis: a metric approach 1909
Why do some research articles receive more online attention? 
Reasons for online success as measured with altmetrics

Kim Holmberg\(^1\) & Julia Vainio\(^2\)

\(^1\)kim.j.holmberg@utu.fi
Research Unit for the Sociology of Education, University of Turku (Finland)

\(^2\)jukava@live.com
Research Unit for the Sociology of Education, University of Turku (Finland)

Abstract
Recent altmetrics research has started to investigate the meaning of altmetrics and whether altmetrics could reveal something about the attention or impact connected to research. This research continues this line of investigations and studies reasons for why some research has received significant online attention in one or both of two social media services; Twitter or Mendeley. This research investigated Finnish researchers’ opinions about the reasons for why their research had received significant online attention and if the attention received could reflect scientific or societal impact of their research. Based on the findings it can be stated that the level of online attention received is a sum of many factors and that there are also specific differences between the platforms where the attention has been received. For the articles that had received significant attention on Mendeley the reasons for that attention were more often seen as due to an academic audience, while the situation was reverse on Twitter, with the majority of reasons for the attention being linked to a wider audience. Similar trend could be seen when asked about whether the online attention could reflect scientific or societal impact, although a clear consensus about whether online attention could reflect any type of impact at all could not be reached.

Conference Topic
Altmetrics

Introduction
Altmetrics has been defined as “the study and use of scholarly impact measures based on activity in online tools and environments” (Priem, 2014). The research area thus focuses on investigating what mentions of research products (e.g., scientific articles, datasets, code) on the web and particularly in social media can tell about the impact or influence of the research. While the debate about what is meant with impact and whether altmetrics can reflect impact or should we instead be talking about attention or engagement continues, altmetrics research has focused on investigating what the data can reflect and what it can tell us about the research that have received the online attention. Early altmetrics research focused on investigating possible connections between citations and different altmetrics and discovered certain connections between for instance usage statistics on Mendeley (Mohammadi & Thelwall, 2013), Wikipedia references (Evans & Krauthammer, 2011), and tweets mentioning scientific articles (e.g., Shuai et al., 2012; Thelwall et al., 2013), and the numbers of citations those articles later receive. More recent altmetrics research has moved on to investigate the meaning of altmetrics and whether altmetrics could reveal something about the reach, influence, or engagement connected to research. The present research continues this later line of investigations and studies possible reasons for why some research has received significant online attention in one of two social media services, namely Twitter or Mendeley. With the help of a questionnaire we will investigate Finnish researchers’ own opinions about the possible reasons for why their research have received significant online attention and whether the researchers feel that the attention received could reflect either scientific or societal impact of their research.
**Background**

While early altmetrics research focused on investigating relations between citation counts and different altmetrics (e.g., Thelwall et al., 2013), more recent research have focused increasingly on identifying characteristics of the research articles that may influence the level of online attention the articles receive. Haustein, Costas and Larivière (2015) investigated possible connections between document characteristics (discipline, document type, title length, number of pages and references) and five different altmetrics (Twitter, Facebook, blogs, Google+, and mainstream media). Although the coverage of scientific articles was very low on the selected social media platforms (with the possible exception of Twitter), the authors discovered that “both citations and social media metrics increase with the extent of collaboration and the length of the reference list.” The results also showed how editorials and news items were the most popular scientific document types shared on Twitter. Didegah et al. (2016) conducted a similar study investigating a range of factors associating with citation and altmetric counts of Finnish research articles. The factors investigated were individual collaboration (number of authors), international collaboration (number of different countries that the authors came from), institutional impact (Maximum Mean Normalized Citation Score of the different institutions that the authors were affiliated with), Journal Impact Factor (retrieved from the JCR), journal open accessibility, and field type (OECD field of the article). These were then tested for association between citation counts, number of readers on Mendeley, number of tweets on Twitter, and number of posts on Facebook. The main conclusion was the same as in Haustein et al. (2015), that the factors driving citations are very much different from those that drive altmetrics, emphasizing that altmetrics cannot be considered as a replacement for citations, but instead possibly as a complement. In Didegah et al. (2016) most of the tested factors were significantly determining the altmetrics, either associating with an increase or a decrease of the altmetrics. The differences between some of the results in Haustein et al. (2015) and Didegah et al. (2016) also demonstrate the uncertainty present in altmetrics, as factors such as timing of the study and data quality may influence research results. Thus more research and new types of approaches are needed in order to fully understand what different altmetrics may reflect and why some research receive more online attention compared to others.

While most of previous altmetrics research has utilized mainly quantitative research methods, qualitative approaches have been almost non-existent. Qualitative content analysis, for instance, could bring new viewpoints into altmetrics research. Qualitative content analysis can be used as a separate research method as well as a theoretical context provider that can be subjected to several different forms of analysis (Tuomi & Sarajärvi, 2009, 91). Qualitative content analysis can be described as a subjective method of research that enables the interpretation of data through a series of classifications and through the underlying themes of content (Hsieh & Shannon, 2005; Zhang & Wildemuth, 2009). The purpose of qualitative research is not to produce quantifiable information or to provide statistical significance for the data. Instead, it can be used to help us understand the surrounding reality and to enable us to classify it into different codes of conduct, appearing themes and emerging categories (Zhang & Wildemuth, 2009). The present research will take a more qualitative approach to investigate the possible reasons why some research receives significant online attention on Twitter and on Mendeley. Whereas Mendeley is a social reference manager owned by Elsevier, Twitter is a microblog service with over 300 million active monthly users. Mendeley is mostly used by academics in their work (Mohammadi, Thelwall, Haustein, & Larivère, 2015), while Twitter is used for much wider variety of reasons and purposes (e.g., Ke, Ahn, & Sugimoto, 2016; Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011; Semertzidis, Pitoura, & Tsaparas, 2013; Uddin, Imran, & Sajjad, 2014). Both of the selected altmetrics data sources, Twitter and Mendeley, are clearly different in both their purpose and user base,
which will give this research an interesting starting point to examine differences in the possible reasons for the accumulated attention.

**Data and methods**

The goal of this research was to investigate what authors of scientific articles that had received significant attention on either Twitter or Mendeley thought had contributed to the received attention and what they felt that attention could tell, if anything, about the impact of that research.

The overall research design was to create a questionnaire that would demonstrate how the participants view the success of their research articles in two data sources of altmetrics; Twitter and Mendeley. Researchers who during 2012 to 2014 had an affiliation to a Finnish research institute or a university and whose article or articles scored in the top one percent of the most shared articles on Twitter or most saved articles on Mendeley were chosen to participate in the survey. All scientific articles with at least one author with a Finnish affiliation were retrieved from the national VIRTAA research publication database. Using DOIs the articles were matched with altmetrics data aggregated and provided for this research by Altmetric LLP (https://www.altmetric.com/). The articles were then sorted in descending order based on how many times they were tweeted or how often they had been saved on Mendeley. Of these only the articles that made the top 1% percent on either platform were chosen for this study. This resulted in a total of 109 articles that were frequently tweeted and 102 articles that were often saved on Mendeley. These articles had a total of 465 authors with a Finnish affiliation, which made the set of participants that were contacted. Of these 170 researchers (36.6 percent of all researchers included in the data) answered the questionnaire. Due to missing or closed email addresses, 69 researchers (14.8 percent) could not be reached at all. In some cases, the researcher had published an article that faired in more than one online service. If this happened, their answer was registered to the data multiple times: the same answer for both individual online services. However, in the study, each participant was only observed once, i.e. each participant was counted as a single observation.

The questionnaire was conducted as an email survey and the content of it was both in English and in Finnish. Of all the participants, 10 percent (20 people) replied in English. The participants were introduced to two questions regarding the online attention their article received, questions that also function as the overall research questions of this research:

1. **Which factors do you think have influenced the attention this paper has received on the above-mentioned platforms?**
2. **What do you think this online attention tells about your paper and its societal or scientific impact?**

Unlike in altmetrics research in general, the aim of this study was to utilize mainly qualitative methods in assessing the research questions. The data analysis based on the answers received first started with a thorough read-through of the answers. After forming a cohesive view on the data, the text was then submitted to a further, more specific inspection. The data were scanned for words, expressions or sentences that seemed to appear regularly from one answer to another. As instructed by Hsieh and Shannon (2005), the impressions the data left for the researcher were also committed to paper and used in the coding of the answers. Finally, the whole data were coded to different word or phrase categories. The process of coding was performed by using inductive analysis, where the coding takes place by creating new categories when necessary whilst reading through the data instead of submitting the data to a set of fixed categories. By using the means of inductive content analysis, we were able to
perceive the forming of different categories based on the material (Zhang & Wildemuth, 2009; Tuomi & Sarajärvi, 2009, 96).

Defining the unit of coding is one of the most important things to take into consideration (Weber, 1990). In qualitative research, the definition of the coded unit is often times more abstract than just an individual word, a sentence or a paragraph (Zhang & Wildemuth, 2009). Therefore in our material, the categories were formed mainly based on individual words, but with the context of the word taken into consideration as well. Depending on the context and the way the respondents answered the proposed question, some of the answers could have been coded into several different categories (Tesch, 1990), as was done with our data. There were some overlapping categories, as the inductive content analysis and the possibility of coding answers into several different categories allowed for this to happen. However, some of these overlapping categories were later merged together wherever possible and their possible negative effects on analysis have been identified during the stages of analysis and discussed by the authors. Towards the end of the process of analysis we strove towards discovering a larger context for the categories identified – main themes that might serve best in describing the opinions of researchers about the success of their research in the analysed online services (Hsieh & Shannon, 2005).

Of the answers from the 170 respondents, 109 answers concerned an article or articles that were among the top one percent of most tweeted research articles on Twitter and 95 concerned an article or articles that were among the top one percent of the most saved references on Mendeley. As some articles appeared in the top one percent on both Twitter and on Mendeley and these answers were inserted to both platforms, the total number of answers ended up being higher than 170.

In terms of responses by online service, for the articles that were among the top one percent of articles that were most frequently saved on Mendeley researchers from the Natural Sciences answered the questionnaire most frequently – 50.5 percent (48 persons) in total. Researchers from Engineering and technology answered the questionnaire with an 34.7 percent share (33 persons), while the share of Medical and Health Science researchers that responded to the questionnaire regarding success in Mendeley, was only 10.5 percent (10 persons) responding and with Social Sciences researchers amounting to about 4.2 percent (4 persons). With 49.5 percent (54 persons) of the answers to Twitter Medical and Health Sciences researchers provided the most answers. Researchers from Engineering and Technology answered the questionnaire with a share of 28.4 percent (31 persons), while the share of answers from Natural Sciences researchers was only 20 respondents (18.3 percent) of the possible 109 people responded to the questionnaire regarding success on Twitter. Respondents from Social Sciences amounted to about 3.7 percent (4 persons).

Results

As there was no prevailing theoretical frame of reference that would have provided us with hypotheses over the answers, we settled on utilizing the previously introduced method of inductive content analysis by Hsieh and Shannon (2005) to analyse the answers to the questionnaire. With inductive content analysis we were able to identify several reoccurring factors that were mentioned to be behind the online success (question 1) or that were thought to reflect social and scientific impact (question 2) of the paper in question. The overarching themes that were built by further grouping the categories in question 1 were constructed based on what sort of audience the researchers might have envisioned when they described the reasons for the success of their article. After the first stage of analysis we were able to interpret the data and identify two dominant overarching themes – attention from an academic audience and attention from a wider audience beyond academia. The chosen themes, academic and wider audience, were created based on what type of audience the
response could be thought to intend. In other words, in the creation of themes the audience was taken into account. For instance, when considering the reasons stated for the popularity or gained attention the category for “respected publication channel” was coded to be more important for an academic audience, rather than for a wider audience beyond academia. The category “timeliness” on the other hand was coded to be mainly appealing for a wider audience. There were some overlapping between the categories (e.g., timeliness – novelty), but these were not merged together because of the subtle nuances in the meaning and intention of the responses. Behind this decision were the perceived motivations to share the article on Twitter or to save it on Mendeley. It was presumed that an academic audience has differing motivations to share or save the article than a wider, mainly non-academic audience would have.

Table 1 shows the identified themes from the responses to the first question grouped according to the overarching themes (academic audience – wider audience). The table is sorted descending according to Mendeley categories. Certain categories were clearly more popular for articles that had received significant attention in both Twitter and Mendeley, such as timeliness (Twitter 20 mentions, Mendeley 19 mentions), respected publication channel (T 20, M 23), popularizable topic (T 16, M 12), activity of the authors in disseminating information about the article (T 8, M 8), and review/meta-analysis article (T 7, M 13). As can be seen in Table 1, responses connected to articles that had received significant attention on Twitter emphasized more personal attributes and the role of communication, while responses connected to articles saved on Mendeley reflected a more academic and analytical approach to the factors leading to the received online attention. Where almost 68 percent of the responses for Mendeley were classified as reasons related to an academic audience, for Twitter about 60 percent of the responses were classified as attributes appreciated by a wider audience. For responses connected to Twitter five topics were clearly emphasized as reasons for the received attention; emotionally engaging topic (30 mentions on Twitter, 0 mentions on Mendeley), respected publication channel (T 20, M 23), timeliness of the topic (T 20, M 19), novelty of the topic (T 17, M 6), and popularizable topic (T 16, M 12). For Mendeley the responses were more evenly spread over more topics. There were a total of eight categories that received over ten mentions in the responses; significant research results (40 mentions on Mendeley, 8 mentions on Twitter), respected publication forum (M 23, T 20), timeliness (M 19, T 20), new research method (M 18, T 2), review/meta-analysis article (M 13, T 7), popularizable topic (M 12, T 16), top researchers as authors (M 12, T 2), and interesting topic (M 12, T 0).
Table 1. Factors influencing the attention that top research articles had received on Twitter and Mendeley (A = Academic audience, W = Wider audience).

<table>
<thead>
<tr>
<th>Theme Category</th>
<th>Twitter (n)</th>
<th>Mendeley (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A significant research result</td>
<td>8</td>
<td>40</td>
</tr>
<tr>
<td>A respected publication channel</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td>A new research method</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>A review article / meta analysis</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>A top researchers among the authors</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>A activity of the authors (in disseminating the results)</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>A surprising result</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>A well known research area</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>A conflicting results</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>A highly cited</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>A international collaboration</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>A connection to the Nobel prize</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>A multidisciplinarity</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>A access to data</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>A basic research</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A networks of the authors</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>A consortium research</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>A wrong interpretations</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>76</strong></td>
<td><strong>158</strong></td>
</tr>
<tr>
<td><strong>%</strong></td>
<td><strong>40.2%</strong></td>
<td><strong>67.8%</strong></td>
</tr>
<tr>
<td>W timeliness</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>W popularizable or personal topic</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>W interesting topic</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>W engaging to wider audience</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>W novelty</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>W visibility in media</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>W open access</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>W press release by the institution / university</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>W title</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>W figures</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>W easily understood</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>W emotionally engaging topic</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>W applicability in practice</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>W societal relevance</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>113</strong></td>
<td><strong>75</strong></td>
</tr>
<tr>
<td><strong>%</strong></td>
<td><strong>59.8%</strong></td>
<td><strong>32.2%</strong></td>
</tr>
<tr>
<td>Total</td>
<td><strong>189</strong></td>
<td><strong>233</strong></td>
</tr>
<tr>
<td><strong>%</strong></td>
<td><strong>100 %</strong></td>
<td><strong>100 %</strong></td>
</tr>
</tbody>
</table>
In the responses to the second question there were again some identified topics common to both platforms. The responses were again grouped into overarching themes based on what the online attention was thought to reflect or indicate and some clear differences between the platforms in regard to the second question emerged. Figure 1 shows that online attention was in many cases thought to reflect scientific impact of the article, especially on Mendeley (Twitter 12 mentions, Mendeley 45 mentions). Authors of the articles that had received significant attention on Twitter emphasized their opinion that the online attention was a reflection of some type of impact other than scientific impact (33 respondents), but almost as many respondents (24) stated that they did not believe that attention on Twitter reflected any type of impact. Thus, no clear consensus about the impact could be reached among the authors of the articles that had received significant attention on Twitter. Same, although to a lesser degree could be seen for Mendeley, with 45 responses stating that the online attention is a reflection of scientific impact, 14 responses claiming it to be a reflection of societal impact, and 12 respondents claiming that online attention does not reflect any type of impact. On the other hand, some authors of articles that had received significant attention on Mendeley stated that the online attention correlated with the number of citations their research had received and that it reflected how the scientific community was interested in the research. In addition, other aspects of the research or of its authors (Twitter 22 mentions, Mendeley 39 mentions) and quality of the research (T 18, M 28) were also quite often mentioned as being reflected in the online attention received.

![Figure 1. Respondents’ opinions about what online attention on Twitter and Mendeley tells about the article’s impact, about the research or of other attributes of the article or its authors](image)

**Discussion**

The overall aim of this study was to investigate Finnish researchers’ opinions about the reasons why their research had received significant attention online on either Twitter or Mendeley or on both. Some shared attributes of the research or the publication appeared repeatedly for both Twitter and Mendeley. Especially high quality of the publication channel, timeliness of the publication, and personal connection to the research topic were seen as attributes that influenced the online attention on both platforms. In addition, depending on the
platform, top researchers as authors, emotionally appealing subject, and the novelty of the research topic were also often mentioned as reasons for the received attention. Based on the findings it can be stated that the level of online attention received is a sum of many factors and that there are specific differences between the platforms where the attention has been received. The categories were grouped according to whether the researchers thought that the online attention was due to attention received from an academic audience or from a wider more general audience. From this grouping it was apparent that for the articles that had received significant attention on Mendeley the reasons for the attention were more often due to an academic audience, while the situation was reverse on Twitter, with the majority of reasons for the attention being linked to a wider audience beyond academia. These platform specific differences in the reasons for the attention can at least partially be explained with both differing purposes and user types between the platforms. While Mendeley is profiled clearly as a tool to conduct research, the user base of Twitter is much more varied.

The divide between an academic audience and a wider audience was also apparent in the answers to the second question about the researchers own thoughts whether this attention could reflect or tell something about scientific or societal impact of their research. The impact of the articles frequently saved on Mendeley was described to be more connected to a scientific impact, while the articles that were frequently tweeted were thought to reflect the attention from a wider audience. But some disagreement could be seen when the researchers assessed the scientific and the societal impact of their article. The abstract nature of the term impact and the subjective opinions attached to the term made some respondents specify that the online attention gives the research and the research topic visibility, but not impact. On the other hand, in some answers the articles were seen to have some general impact and influence on societal debate. But several respondents stated clearly that they do not believe that the amount of attention on Twitter would tell anything about any possible impact of that research.

As it would appear that all the respondents recognized both of the platforms, it is possible that their answers have to some degree been influenced by their personal perception of the platforms and of their users. This could be counted for by repeating the study without revealing on which specific online platforms the articles have received significant attention. Although there were clear differences in the answers according to the platform, it is unclear how much the answers were influenced by researchers’ own perception of the type of people that are using the platforms. There was a lack of clear consensus about whether the online attention received could reflect scientific or societal impact. Opinions that the online attention contributed positively to impact in general prevailed, although some described the online attention as mere buzz. Traditional approaches to measure impact, such as citation counts, were seen among the respondents as the most important means to measure impact.

The results did, however, show a consensus among the researchers about the importance of breaking out of the ivory towers and communicating research to a wider audience. The answers often repeated the researchers’ view that the online attention received demonstrated how the researchers had fulfilled their societal duty and also how they could bring scientific viewpoints into policy making. More research is, nevertheless, needed to fully understand what the online attention some research receive could tell about societal impact of research or how science has had an impact on policy making.
References


Can Twitter Increase the Visibility of Chinese Publications?

Fei Shu¹ and Stefanie Haustein²

¹fei.shu@mail.mcgill.ca
McGill University, Montreal (Canada)

²stefanie.haustein@umontreal.ca
Université de Montréal, Montreal (Canada)

Abstract
The purpose of this study is to investigate whether diffusion on social media can help to improve the international visibility of Chinese papers and thus increase citation impact. After analyzing 160,233 Chinese articles published in 2012 as well as their tweets and citations received, the results indicate that, tweeted Chinese papers, published in the same year as the same journal, received around 20% more citations than Chinese papers not mentioned on twitter. The citation advantage of tweeted Chinese paper is also found within various disciplines and by different citing source countries.

Conference Topic
Twitter, citation analysis, China, altmetrics, scientific impact

Introduction
With the significant development of the Chinese economy, Chinese scientific activity is experiencing a period of similar growth. China¹ has become the second largest source country in terms of its share of international scientific production since 2009 (ISTIC, 2010, 2015). A number of studies attempt to characterize Chinese research achievement and evaluate China’s research performance regarding almost all bibliometric topics such as collaboration network (Niu & Qiu, 2014; Wang, Wu, & Pan, 2014), academic journals (Ding, Zheng, & Wu, 2012), university ranking (Fu & Ho, 2013; Zhu, Hassan, Mirza, & Xie, 2014), and international visibility (Ren & Rousseau, 2002; Wu et al., 2004). However, this is not the case in altmetric studies that few researches try to measure the dissemination of China’s research via non-traditional sources (e.g. news report, social media, blogs, reference management tools etc.).

Although China is still isolated from the major international social media such as Twitter, Facebook, Google¹, China’s research has been mentioned in these social networks and acquires international attention. Unfortunately, so far no current study addresses whether mentions in social media can increase the citation impact of China’s research.

Although the altmetric research investigates almost all social media as tools measuring the influence of research activities, Twitter has been in the centre of altmetric research because of its coverage (Thelwall, Haustein, Larivière, & Sugimoto, 2013). Comparing to other social media, Twitter shows the largest signal related to scientific papers among non-academic, general audience social media platforms. About one fifth of recent publications are mentioned on Twitter, which is topped only by Mendeley among currently captured altmetrics (Haustein, Costas, & Larivière, 2015). Although Twitter is not accessible in China², China’s publications are tweeted internationally. These international tweets could improve the international visibility

¹ In this study, China refers to Mainland China, which is the area under the direct jurisdiction of the People's Republic of China (PRC) but excluding Hong Kong and Macau.
² People can occasionally access the Twitter by using the virtual private network (VPN) but it is illegal in China.
of China’s publication. However, no study reveals whether more international visibility may bring more citations to China’s publication or not. In addition, since Chinese scholar cannot use Twitter to promote their publications, the investigation on China’s publications provides a special environment to understand the impact of tweets on citations when the influence of self-tweets is excluded.

Although some studies have associated Twitter exposure with an overall increase of citations (Eysenbach, 2011; Thelwall et al., 2013), the direct correlations between tweets and citations have been shown to be weak (Costas, Zahedi, & Wouters, 2015; de Winter, 2015; Haustein et al., 2015). Moreover, social media and altmetrics have been heralded as democratizers of science and its reward system, as they potentially overcome the Matthew effect reflected in traditional citation-based indicators. However, low social media visibility of scientific publications from Latin America (Alperin, 2014, 2015) and Iran (Maleki, 2014) suggest that instead of leveling the playing field and contributing to global equality in scholarly communication, social media seem to widen the gap between countries. Despite being ranked second largest source country in terms of the number of publications (ISTIC, 2015), several studies show that Chinese publications still receive low international visibility in terms of the number of citations received (Ren & Liang, 1999; Ren & Rousseau, 2002; Wu et al., 2004). We do not know whether the diffusion of paper on Twitter can increase China’s international visibility and improve its citation impact.

Goals and Objectives
The purpose of this study is to investigate whether diffusion of papers on Twitter can help to increase the visibility of Chinese papers and thus help to close the citation impact gap between China and other countries. It seeks to answer the following questions:

1. Does the overall citation rate of Chinese papers shared on Twitter exceed those of non-tweeted Chinese papers?
2. How do results described above vary by scientific disciplines?
3. How do results described above vary by the source countries of tweets and citations?

Methodology
In this study, the publication and citation data were retrieved from Web of Science (WoS), which includes the Science Citation Index Expanded, the Social Science Citation Index and the Arts and Humanities Citation Index, annually index documents published in about 12,000 journals, covering all areas of research. In this study, all papers published in 2012 were retrieved (N = 1,339,279). Citations to these 2012 articles were counted until the end of 2015, which allows for a citation window of three complete years.

Altmetric.com was chosen as the data source for tweet counts, as it is the most comprehensive source of social media data associated with scientific papers (Robinson-Garcia, Torres-Salinas, Zahedi, & Costas, 2014). Tweet data includes counts collected up to June 2015.

The Digital Object Identifier (DOI) was used as the linkage between WoS and Altmetric.com data. Since the DOIs are not available in all publications, only publications with a DOI as recorded in WoS (N = 1,131,358) were included under investigation.

Based on the dataset of 1,131,358 articles, we identified three groups of Chinese papers published in 2012 for investigation: 160,233 articles have at least one Chinese author (Group 1); the first authors are from China in 145,325 articles (Group 2); and 119,268 articles only have Chinese author(s) (Group 3). In this study, the Chinese author was defined as the author...
with an affiliation to a Chinese institution. The bibliographic information such as journal title, discipline, and citations to these articles were also retrieved from the WoS. As Table 1 indicates, almost 40% of the journals contain tweeted Chinese papers in each group. These journals contribute more than two thirds of the Chinese papers.

Table 1. Number of journals publishing Chinese papers (2012)

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of articles</th>
<th>Number of journals</th>
<th>Number of journals with tweeted articles</th>
<th>Number of articles under investigation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>160,233</td>
<td>6,365</td>
<td>2,508 (39.4%)</td>
<td>109,334 (68.2%)</td>
</tr>
<tr>
<td>2</td>
<td>145,325</td>
<td>6,004</td>
<td>2,485 (41.4%)</td>
<td>98,372 (67.7%)</td>
</tr>
<tr>
<td>3</td>
<td>119,268</td>
<td>5,351</td>
<td>2,372 (44.3%)</td>
<td>80,394 (67.4%)</td>
</tr>
</tbody>
</table>

Since the citation frequencies vary across disciplines, the comparisons between disciplines are difficult. Based on the journals in which these Chinese papers were published, all articles were grouped into 14 major disciplines based on the NSF journal classification system (See Table 2). Since Chinese scholars published very few articles in Arts and Humanities, these two disciplines were excluded from this study.

Table 2. Distribution of Chinese papers (Group 1) by discipline

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Number of tweeted papers</th>
<th>Number of non-tweeted papers</th>
<th>Total</th>
<th>Ratio of Tweeted paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts</td>
<td>1</td>
<td>14</td>
<td>15</td>
<td>6.7%</td>
</tr>
<tr>
<td>Biology</td>
<td>766</td>
<td>7,109</td>
<td>7,875</td>
<td>9.7%</td>
</tr>
<tr>
<td>Professional Fields</td>
<td>300</td>
<td>1,555</td>
<td>1,855</td>
<td>16.2%</td>
</tr>
<tr>
<td>Chemistry</td>
<td>3,344</td>
<td>26,412</td>
<td>29,756</td>
<td>11.2%</td>
</tr>
<tr>
<td>Engineering</td>
<td>1,119</td>
<td>34,246</td>
<td>35,365</td>
<td>3.2%</td>
</tr>
<tr>
<td>Humanities</td>
<td>11</td>
<td>79</td>
<td>90</td>
<td>12.2%</td>
</tr>
<tr>
<td>Mathematics</td>
<td>352</td>
<td>6,451</td>
<td>6,803</td>
<td>5.2%</td>
</tr>
<tr>
<td>Clinical Medicine</td>
<td>5,463</td>
<td>19,543</td>
<td>25,006</td>
<td>21.8%</td>
</tr>
<tr>
<td>Physics</td>
<td>1,537</td>
<td>19,456</td>
<td>20,993</td>
<td>7.3%</td>
</tr>
<tr>
<td>Psychology</td>
<td>147</td>
<td>437</td>
<td>584</td>
<td>25.2%</td>
</tr>
<tr>
<td>Biomedical Research</td>
<td>5,198</td>
<td>15,232</td>
<td>20,430</td>
<td>25.4%</td>
</tr>
<tr>
<td>Health</td>
<td>234</td>
<td>437</td>
<td>671</td>
<td>34.9%</td>
</tr>
<tr>
<td>Earth and Space Science</td>
<td>700</td>
<td>8,759</td>
<td>9,459</td>
<td>7.4%</td>
</tr>
<tr>
<td>Social Science</td>
<td>187</td>
<td>1,144</td>
<td>1,331</td>
<td>14.0%</td>
</tr>
<tr>
<td>Total</td>
<td><strong>19,359</strong></td>
<td><strong>140,874</strong></td>
<td><strong>160,233</strong></td>
<td><strong>12.1%</strong></td>
</tr>
</tbody>
</table>

Three types of analyses were conducted to answer the research questions as above. A pair comparison of average citations received between tweeted Chinese papers and non-tweeted Chinese papers published in the same journal the same year could answer the first research question; the different results from different disciplines will answer the second research question; a further analysis investigating the geographic distribution of these tweets and citations received could answer the third research question.

Journal Comparison

Since the paper’s visibility may be influenced by various factors (e.g. journal is probably a confounding factor of the Twitter impact and citation impact), a comparison of papers published
in the same year in the same journal will provide an experiment by controlling the effect of other factors on the paper’s visibility. In this study, the pair comparison compared the average number of citations to tweeted Chinese papers with the average number of citations to non-tweeted Chinese papers journal by journal. Although more than 2,000 journals contain both tweeted Chinese papers and non-tweeted Chinese papers in each defined group, the distribution of tweets is highly skewed as Figure 1 shows. More than half of these journals only have 1 or 2 tweeted Chinese papers. In order to avoid the effect of outliers, two thresholds were set as below:

- Threshold 1: the qualified journals have at least both 10 tweeted Chinese papers and 10 non-tweeted Chinese papers.
- Threshold 2: the qualified journals have at least both 20 tweeted Chinese papers and 20 non-tweeted Chinese papers.

Eventually, less than 300 journals were selected for the pair comparison as Table 3 indicates. Since the citation data is extremely skewed in the power law distribution, the geometric mean, comparing to the arithmetic mean, is the most precise and accurate indicator for citation-based comparison (Thelwall, 2016). In order to allow the geometric mean to include the uncited articles, 1 is added to the citation counts before calculating the geometric mean and then 1 is subtracted from the result. The shift of 1 is a standard method for calculating the geometric mean of citation data (Thelwall, 2016). In addition, since the citation rate varies among different disciplines, in order to avoid the disciplinary bias, all citations were normalized in this study before counting when the comparison was among different disciplines.

![Figure 1. Distribution of tweets to Chinese papers by journals](image)

**Table 3. Number of journals under investigation by group**

<table>
<thead>
<tr>
<th>Group</th>
<th>No Threshold</th>
<th>Threshold 1</th>
<th>Threshold 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2,508</td>
<td>261</td>
<td>114</td>
</tr>
<tr>
<td>2</td>
<td>2,485</td>
<td>226</td>
<td>96</td>
</tr>
<tr>
<td>3</td>
<td>2,372</td>
<td>167</td>
<td>77</td>
</tr>
</tbody>
</table>
Discipline Comparison

The comparison of citation rate was also conducted among 12 disciplines described above when Arts and Humanities were excluded. All Chinese papers, either tweeted or non-tweeted, were grouped into 12 disciplines based on their journals. To detect general trends and differences between disciplines, we compared, for each discipline, the arithmetic mean of the citations received for tweeted Chinese papers and for non-tweeted Chinese papers. Since few outliers (i.e. highly cited papers) might influence the average number of citations, the analysis was done not only at the paper level but also at the journal level, comparing the mean of citations of 261 journals that was included in Group 1 with Threshold 1 in the journal comparison.

Source Country Comparison

The source countries of all tweets and citations to the 160,233 Chinese papers were retrieved and identified during the data collection. If authors of the citing paper come from different countries, each source country was counted as one when calculating the number of citing source countries. The tweeter country was retrieved from the Altmetric.com who identify it based on the information provided in the twitter account information; unfortunately, the source countries of 18,589 tweets (34.4%) could not be located based on the twitter account information, which is a limitation of this study. For the top 10 source countries in terms of the number of tweets and the number of tweeted Chinese papers, both the normalized geometric mean and the normalized arithmetic mean of the number of citations to Chinese papers were compared. We focused this analysis on group 3 (Chinese author only) in order to avoid the influence of self-tweets because Chinese authors cannot tweet due to the inaccessibility of Twitter in China.

Results

As Table 4 shows, 12.1% of the 160,233 articles with at least one author from a Chinese institution (Group 1) were tweeted and received 53,972 tweets. This percentage is far below the worldwide Twitter coverage of 21.9%. This ratio in Group 2 (11.0%) and Group 3 (9.5%) is even lower, as is to be expected considering the less central role or absence of international co-authors, respectively. Papers in Group 1 have more chance to be tweeted than papers in Group 2 and 3 because international co-authors may promote their co-authored papers on Twitter.

The 160,233 Chinese papers received a total of 1,514,138 citations between 2012 and 2015. As Table 5 shows, the average number of normalized citations received by tweeted Chinese papers is about more than 50% higher than that of non-tweeted Chinese papers received in each group.

### Table 4. Number of tweets received to Chinese papers (2012)

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of articles</th>
<th>Number of tweeted articles</th>
<th>Number of Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>160,233</td>
<td>19,359 (12.1%)</td>
<td>53,972</td>
</tr>
<tr>
<td>2</td>
<td>145,325</td>
<td>15,951 (11.0%)</td>
<td>36,921</td>
</tr>
<tr>
<td>3</td>
<td>119,268</td>
<td>11,319 (9.5%)</td>
<td>22,846</td>
</tr>
</tbody>
</table>

### Table 5. Number of citations received to Chinese papers (2012)

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of citations</th>
<th>Normalized citation rate</th>
<th>Normalized citation rate (tweeted article)</th>
<th>Normalized citation rate (non-tweeted article)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,514,138</td>
<td>1.071</td>
<td>1.638</td>
<td>0.993</td>
</tr>
<tr>
<td>2</td>
<td>1,286,685</td>
<td>1.016</td>
<td>1.441</td>
<td>0.963</td>
</tr>
<tr>
<td>3</td>
<td>988,265</td>
<td>0.939</td>
<td>1.338</td>
<td>0.896</td>
</tr>
</tbody>
</table>

3 There is a possible issue pertaining to the accuracy of the data regarding the source countries that Twitter users can select any countries in their profile.
Journal Comparison

In the journal comparison, when comparing the geometric mean of the number of citations to tweeted Chinese papers with non-tweeted Chinese papers in each journal, we found that tweeted Chinese papers received more citations than non-tweeted Chinese papers in the majority of the journals. As Table 6 indicates, the higher threshold applied, the higher ratio of journals that their tweeted papers have citation advantage.

Table 6. Percentage of journals tweeted paper receiving more citations than non-tweeted paper

<table>
<thead>
<tr>
<th>Group</th>
<th>No Threshold</th>
<th>Threshold 1</th>
<th>Threshold 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>58.9%</td>
<td>67.8%</td>
<td>78.1%</td>
</tr>
<tr>
<td>Group 2</td>
<td>57.8%</td>
<td>68.1%</td>
<td>78.1%</td>
</tr>
<tr>
<td>Group 3</td>
<td>58.1%</td>
<td>68.9%</td>
<td>76.6%</td>
</tr>
</tbody>
</table>

Tweeted Chinese papers received more citations than non-tweeted Chinese papers when comparing both the geometric mean and the arithmetic means in each group and threshold at the journal level. As Table 7, 8 and 9 shows, the tweeted Chinese papers received around 20% more citations than non-tweeted Chinese papers in average if only comparing the arithmetic mean.

Table 7. The comparison of average citations to Chinese papers in Group 1

<table>
<thead>
<tr>
<th></th>
<th>Threshold 1</th>
<th>Threshold 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Geometric mean</td>
<td>Arithmetic mean</td>
</tr>
<tr>
<td>Tweeted paper</td>
<td>9.507</td>
<td>15.584</td>
</tr>
<tr>
<td>Citation</td>
<td>14.5%</td>
<td>22.4%</td>
</tr>
</tbody>
</table>

Table 8. The comparison of average citations to Chinese papers in Group 2

<table>
<thead>
<tr>
<th></th>
<th>Threshold 1</th>
<th>Threshold 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Geometric mean</td>
<td>Arithmetic mean</td>
</tr>
<tr>
<td>Tweeted paper</td>
<td>9.352</td>
<td>15.520</td>
</tr>
<tr>
<td>Citation</td>
<td>15.1%</td>
<td>22.0%</td>
</tr>
</tbody>
</table>

Table 9. The comparison of average citations to Chinese papers in Group 3

<table>
<thead>
<tr>
<th></th>
<th>Threshold 1</th>
<th>Threshold 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Geometric mean</td>
<td>Arithmetic mean</td>
</tr>
<tr>
<td>Tweeted paper</td>
<td>8.952</td>
<td>15.408</td>
</tr>
<tr>
<td>Non-tweeted</td>
<td>7.782</td>
<td>12.758</td>
</tr>
<tr>
<td>Citation</td>
<td>15.0%</td>
<td>20.8%</td>
</tr>
</tbody>
</table>

Discipline Comparison

All 160,233 articles were grouped into 14 disciplines as Table 2 described above. Arts and Humanities were excluded from the data analysis because of the data size. When comparing the arithmetic mean of citations at the paper level, as Table 10 indicates, we found that the tweeted Chinese papers received more citations than non-tweeted Chinese papers in all disciplines except for Professional fields and Mathematics that the tweeted paper received less citations than non-tweeted paper. On the other hand, tweeted Chinese papers had huge citation advantage in some disciplines such as Engineering (251.0%), Physics (143.8%) and Chemistry (76.1%).

15
As we discussed above, in order to avoid the effect of the few highly cited papers on the mean of the citations, we made another comparison within the 261 journals (Group 1 with Threshold 1). 4 out of 12 disciplines were excluded because they contained less than 2 journals. As Table 10 shows, Tweeted Chinese papers had citation advantage compared to non-tweeted Chinese papers in all 8 disciplines. Tweeted Chinese papers in Mathematics received 32.3% more citations than non-tweeted Chinese papers when the effect of outliers was removed; Tweeted Chinese papers still had citation advantage in Engineering (25.2%), Physics (64.2%) and Chemistry (20.7%); but the difference observed between tweeted and non-tweeted papers in these three disciplines was reduced as compared to the paper level analysis.

Table 10. The impact of tweets on citation by discipline

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Tweeted</th>
<th>Non-tweeted</th>
<th>Citation advantage</th>
<th>Tweeted</th>
<th>Non-tweeted</th>
<th>Citation advantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td>11.280</td>
<td>7.806</td>
<td>44.5%</td>
<td>12.547</td>
<td>10.150</td>
<td>23.6%</td>
</tr>
<tr>
<td>Biomedical Research</td>
<td>12.413</td>
<td>9.050</td>
<td>37.2%</td>
<td>14.212</td>
<td>11.495</td>
<td>23.6%</td>
</tr>
<tr>
<td>Chemistry</td>
<td>21.369</td>
<td>12.134</td>
<td>76.1%</td>
<td>20.436</td>
<td>16.935</td>
<td>20.7%</td>
</tr>
<tr>
<td>Clinical Medicine</td>
<td>9.109</td>
<td>7.422</td>
<td>22.7%</td>
<td>12.352</td>
<td>11.094</td>
<td>11.3%</td>
</tr>
<tr>
<td>Earth and Space Science</td>
<td>9.265</td>
<td>9.198</td>
<td>0.7%</td>
<td>15.117</td>
<td>11.293</td>
<td>33.9%</td>
</tr>
<tr>
<td>Engineering</td>
<td>28.435</td>
<td>8.101</td>
<td>251.0%</td>
<td>33.485</td>
<td>26.751</td>
<td>25.2%</td>
</tr>
<tr>
<td>Health</td>
<td>5.971</td>
<td>4.824</td>
<td>23.8%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Mathematics</td>
<td>2.750</td>
<td>3.899</td>
<td>-29.5%</td>
<td>3.267</td>
<td>2.468</td>
<td>32.3%</td>
</tr>
<tr>
<td>Physics</td>
<td>14.048</td>
<td>5.763</td>
<td>143.8%</td>
<td>16.765</td>
<td>10.212</td>
<td>64.2%</td>
</tr>
<tr>
<td>Professional Fields</td>
<td>5.963</td>
<td>6.304</td>
<td>-5.4%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Psychology</td>
<td>5.935</td>
<td>5.503</td>
<td>7.9%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Social Science</td>
<td>6.403</td>
<td>6.148</td>
<td>4.1%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Source Country Comparison

The 160,233 Chinese papers published in 2012 were cited by scholars from 198 countries, and received tweets from 154 countries. China itself contributed 47.0% of total citations (720,816/1,533,582) to non-tweeted Chinese papers while only 29.3% of the total citations (158,062/538,856) to tweeted Chinese papers were self-citations (i.e. citations received come from a Chinese paper) in terms of the source citing country. When self-citations were excluded, as Table 11 shows, the citation advantage was found in more journals in each group and threshold comparing the results as shown in Table 6. It means that tweets may bring more international citations to Chinese papers.

Table 11. Percentage of journals tweeted paper receiving more citations than non-tweeted paper when between including and excluding the self-citations

<table>
<thead>
<tr>
<th>Group</th>
<th>No Threshold</th>
<th>Threshold 1</th>
<th>Threshold 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>58.9%</td>
<td>67.8%</td>
<td>78.1%</td>
</tr>
<tr>
<td>Group 1 (excluding self-citation)</td>
<td>60.1%</td>
<td>74.7%</td>
<td>82.5%</td>
</tr>
<tr>
<td>Group 2</td>
<td>57.8%</td>
<td>68.1%</td>
<td>78.1%</td>
</tr>
<tr>
<td>Group 2 (excluding self-citation)</td>
<td>57.8%</td>
<td>75.2%</td>
<td>87.5%</td>
</tr>
<tr>
<td>Group 3</td>
<td>58.1%</td>
<td>68.9%</td>
<td>76.6%</td>
</tr>
<tr>
<td>Group 3 (excluding self-citation)</td>
<td>58.8%</td>
<td>73.1%</td>
<td>80.5%</td>
</tr>
</tbody>
</table>

In addition, the source country of tweets has the influence on the citation advantage of tweeted Chinese paper. As Table 12 shows, nine countries appear in both lists of top 10 source countries in terms of the number of tweets and the number of tweeted Chinese papers, United States,
United Kingdom and Australia are ranked top 3 in both list. Although Chinese papers received more tweets from United States, these tweets did not bring any citations advantage comparing to tweets from other source countries. Both the geometric mean and the arithmetic mean of normalized citations to these papers tweeted by American are below the average. On the other hand, Chinese papers received more citations when they were tweeted by German or Japanese.

Table 12. The average citations received from the top 10 source countries in terms of the number of tweets and the number of tweeted Chinese paper (Group 3)

<table>
<thead>
<tr>
<th>Country</th>
<th>Order by number of tweets</th>
<th>Order by number of tweeted Chinese paper</th>
<th>Normalized Geometric mean</th>
<th>Normalized Arithmetic mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>1</td>
<td>1</td>
<td>0.933</td>
<td>1.384</td>
</tr>
<tr>
<td>UK</td>
<td>2</td>
<td>2</td>
<td>0.874</td>
<td>1.192</td>
</tr>
<tr>
<td>Australia</td>
<td>3</td>
<td>3</td>
<td>1.175</td>
<td>1.656</td>
</tr>
<tr>
<td>Argentina</td>
<td>4</td>
<td>4</td>
<td>0.757</td>
<td>1.026</td>
</tr>
<tr>
<td>Canada</td>
<td>5</td>
<td>6</td>
<td>1.102</td>
<td>1.505</td>
</tr>
<tr>
<td>Germany</td>
<td>6</td>
<td>5</td>
<td>1.692</td>
<td>2.504</td>
</tr>
<tr>
<td>Japan</td>
<td>7</td>
<td>7</td>
<td>1.529</td>
<td>2.267</td>
</tr>
<tr>
<td>France</td>
<td>8</td>
<td>8</td>
<td>0.943</td>
<td>1.235</td>
</tr>
<tr>
<td>Spain</td>
<td>9</td>
<td>10</td>
<td>1.118</td>
<td>1.550</td>
</tr>
<tr>
<td>Netherlands</td>
<td>10</td>
<td>11</td>
<td>0.718</td>
<td>0.929</td>
</tr>
<tr>
<td>Ghana</td>
<td>11</td>
<td>9</td>
<td>0.933</td>
<td>1.015</td>
</tr>
<tr>
<td>No Location</td>
<td></td>
<td></td>
<td>0.993</td>
<td>1.433</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>1.002</td>
<td>1.426</td>
</tr>
</tbody>
</table>

Conclusion
In this study, after investigating the tweets and citations to 160,233 Chinese papers published in 2012, we determined the differences on average citations received between tweeted Chinese papers and non-tweeted Chinese papers. The results of this study show tweeted Chinese papers received more citations than non-tweeted Chinese papers:

- From the same journal the same published year, tweeted Chinese papers received about 20% more citations than non-tweeted Chinese papers.
- When the self-citations were excluded, the citation advantage of tweeted Chines paper is bigger,
- The citation advantage of tweeted Chinese paper varies by scientific disciplines.

Since both the tweet data and the citation data were collected within the same 3-year period, we cannot conclude the causation between the tweets and citations received based on the results but indicate the correlation between whether Chinese papers were tweeted or not and the number of citations they received. In addition, there are also some other limitations in this study. The source countries of about one third of tweets could not be located; the co-authorship of citing articles inflates the number of citations when counting them by countries. In order to understand more detailed impact of tweets on citations to Chinese papers, the content of tweets as well as their relationship with the citations to Chinese papers need to be investigated for further analysis.
Acknowledgments
This study is supported by iFellows Doctoral Scholarship provided by the Andrew W. Mellon Foundation and the Doctoral Research Scholarship provided by the Fonds de recherche société et culture Québec (FRQSC).

References


Thelwall, M. (2016). The precision of the arithmetic mean, geometric mean and percentiles for citation data: An experimental simulation modelling approach. *Journal of Informetrics, 10*(1), 110-123. doi: [http://dx.doi.org/10.1016/j.joi.2015.12.001](http://dx.doi.org/10.1016/j.joi.2015.12.001)


Exploratory Analysis on the Construction of Journal Evaluation Model Using Altmetrics Indices

Zhang Yang¹ and Lang Linfang ²

¹ zhyang2@mail.sysu.edu.cn
Sun Yat-Sen University, Guangzhou (China)

² 1932604886@qq.com
Sun Yat-Sen University, Guangzhou (China)

Abstract
The research of Altmetrics is becoming one of the most interesting hot spots in the evaluation science with the use of network data in research study. The Altmetrics study has been changing from exploratory research in the preliminary stage to argumentative research, nowadays. This paper analyzed the differences and interrelationships between citation and Altmetrics by using journals of Information Science & Library Science. After that, the star graph was drawn to visualize the relative size of citation and Altmetrics. At last, the feasibility issue of using two types of indicators in journal evaluation was discussed. The results indicated that the citation data and Altmetrics data didn't confirm to normal distribution. There were not so many journals within high impact factor division, but a lot of journals belonged to low impact factor division. The data distribution revealed that both citation and Altmetrics all displayed the long tail trend. Diffidence was demonstrated between citation indices and Altmetrics indices. The result of correlation analysis also showed a difference within different Altmetrics data. The academic quality, disciplinary difference, platform properly and other factors will work on the visibility and attention to journals on the network. If citation data and Altmetrics data were both used at the same time in journal evaluation, a more comprehensive outcome could be obtained.

Conference Topic
Altmetrics, Indicators, Scientometrics, Bibliometrics

Introduction
A series of journal evaluation indicators are promoted after the concept of Journal Impact Factor presented by Garfield. The Journal Citation Reports (JCR) published by Thomson Reuters provide Journal Impact Factor, Immediacy Index, Normalized Eigenfactor, Citable Items and Article Influence Score. Jorge Hirsch (2007) puts forward the concept of H index, then H index was introduced into the assessment of journals and was developed by Google Scholar. A series of indicators based on H index are promoted by Google Scholar (https://scholar.google.com.hk/intl/zh-CN/scholar/metrics.html#metrics) such as h-core, h-median, h5-index, h5-core and h5-median. Scopus database put forward SNIP and SJR. A number of evaluation indices are used in practice, but we found that most of the indices are based on citation though different scholars and agencies present a variety of evaluation indicators. The journal evaluation methods based on citation expose a series of questions: citation delay, citation difference between subjects, and disadvantage for young scholars with the deepening research. Nowadays, some of scholars try to apply network data in evaluation study in order to explore the feasibility issues of using network data in practice. Altmetrics is entering the field of evaluation science and becoming one of the research hot spots.
**Literature Review**

Altmetrics is short for “Alternative metric” and was studied in the year of 2008. The study originates that Taraborelli D (2008) suspects the issue of using JIF in journal evaluation and propose concepts of Distributed Scientific Evaluation and Alternative Peer Review Models. Priem and Hemminger (2010) propose Scientometrics 2.0 and make a multi-angle analysis of influence in social network data. The official website of Altmetrics was established in October 2010 and published the declaration of Altmetrics: A Manifesto (Priem, Taraborelli & Groth, 2010), symbolizing the establishment of Altmetrics.

From development in 2008 to the present, the network data available for Altmetrics research is abundant, including social media data from Facebook, microblogging data from Twitter, online reference manager data from Mendeley or CiteUlike, expert opinion data from F1000 as well as the comprehensive data from Altmetrics.com. What characteristics do the different source data have? Are there some relations and differences in citation and Altmetrics? In order to work out these questions, a number of researchers carried out exploratory research. Bornmann & Haunschild (2015) analyze users’ preferences for different label literature on F1000 among Mendeley readers. Mohammad etc. (Mohammadi, Thelwall & Haustein, 2015) take five discipline journals as study object, then the authors analyse literature coverage on Mendeley and user structure. Haustein et al apply 1.8 million journal literature included in WOS database as study object and explore the visibility of different type literature on the social media platform (Haustein, Costas & Larivière, 2014). Peters etc. select 200 articles from 50 the most commonly used journals in emergency medicine and find that there is a moderate correlation between citation and Altmetrics (Peters, Kraker & Lex, 2016). Liu etc. visualize the relationship of different evaluation indicators (Liu, Xu, Wu, Chen, & Guo, 2013).

Though Altmetrics originates from the paper level measurement for capturing view quantity, comment count, storage data on social media, online reference manager as well as export recommendation system. Nowadays, the Altmetrics data break through paper level evaluation and begins to explore the application of disciplinary analysis, journal evaluation and institutional evaluation. Chen etc. work out the multidimensional scale analysis on 18 indicators of the humanities and social sciences literature in Taiwan, which including domestic influence index, international influence index, webometrics index and social influence index (Chen, Tang, & Wang, 2015). The results combined with regression analysis and discriminant analysis show that humanities and social sciences has different interpretations in four dimensions. Moed & Halevi (2015) summary the advantages and disadvantages of six widely used indices of citation data, usage data, Altmetrics, patent data, economic data and network data, then they put forward the concept of multidimensional matrix. There are some basic literatures about using Altmetrics data to evaluate journals. Haustein etc. (Haustein, & Siebenlist, 2011) calculate Usage Ratio, Usage Diffusion, Article Usage Intensity and Journal Usage Intensity for journal evaluation by using the Altmetrics data of social bookmarking. Zahedi etc. (Zahedi, Costas, & Wouters, 2015) compare the ability to distinguish the highly cited WOS publication between Mendeley readship counts and citation impact of journal (JCS). Nuredini and Peters (Nuredini, & Peters, 2016) make the case study in 30 journals from Economic and Business studies by using the social media metrics data from Altmetrics.com. The most commonly used Altmetrics data sources for EBS-journal are Mendeley, Twitter and News data. The correlation between Altmetrics score and citation is 0.299 at the paper level, but the coefficient between them is 0.614 at the journal level. In this research, the difference and the relationship between citation data and Altmetrics data will be discussed firstly. Then a star graph will be drawn to visualize the relative size of citation indices and Altmetrics indices. Finally, the study try to construct the comprehensive journal evaluation model including citation data and Altmetrics data by the area of star graph.

The main research questions are as follows:
(1) What kind of data distribution do the citation data and Altmetrics data have?
(2) What kind of relationship do they have between citation and Altmetrics?
(3) How to use the citation dimension data and Altmetrics data to evaluate journal in practice?

Methodology
This study was restricted in the journals of Information Science & Library Science and three citation indices, including Journal Impact Factor, Immediately Index and Eigenfactor Score were retrieved from JCR published by Thomson Reuters database. Three types of Altmetrics data were obtained from Altmetrics.com website. Later, the calculation method of Journal impact factor was drawn on because the data retrieved from Altmetrics.com is based on the paper level and journal level data has to be converted. These three Altmetrics indicators are ReaderIF which represents the average number of journal readers, PostIF which indicates the average number of posts from Twitter of the journal and ScoreIF which represents the average score of the journal on the website of Altmetrics.com. The computation progress is as follows: Firstly, we figured out the total number of reader counts, comment of Twitter and score coming from Altmetrics.com for two years. Then the two years summation of reader counts, comments of Twitter and scores of Altmetrics.com was divided by the amount of 2 years articles. At last, the data reflecting the journal level was acquired. Based on the following reasons, the calculation method of Journal Impact Factor is chose: (1) Although, the social media platform with property of generating user counts instantly, this research referenced two years time window and calculated the total amount Altmetrics of the article published in the previous two years for the reason that journals have a steady number of users, concerns, or comments on social media. The average of Altmetrics data of papers published two years ago of one journal was applied to represent the evaluation index of the journal on social media, because of the paper level data only obtained from the Altmetrics.com. It is assumed that most of the literature published in a journal and received a large number of users or comments under the Altmetrics system should enable the journal obtain a higer value of ReaderIF, PostIF or ScoreIF, which reflects the journal high social concern. In addition, if a small part of the literature of a journal obtained a large number of users or comments, it could also lead the journal to obtain a high value of ReaderIF, PostIF or ScoreIF. This kind of star literature can also bring certain social influence to the journal, thus increasing the number of followers in the social media environment. Therefore, either the most of paper received extensive attention in social media or only a part of paper gained attention will make the publishing journals gain a higer Altmetrics impact factor, reflecting the high impact of journal on social media network. It is theoretically feasible and effective to use ReaderIF, PostIF, and ScoreIF as the Altmetrics evaluation index for journal. That is done in the following analysis using empirical analysis to verify this.

First of all, the literature collection published in 2012 and 2013 were gained form Scopus database for every journal. The use of Scopus database based on the following two reasons: (1) articles in Scopus database can be traced back to 1823 from the perspective of literature collection time. The Scopus database has a large coverage of papers from the Asia Pacific region at the perspective of regional literature coverage. (2) DOI of paper is available in Scopus database compared WOS database, so it's really convenient to collect Altmetrics data by using article DOI. Journal ISSN was used to collect journal paper in Scopus database. There were 87 journals in JCR, however, there were a few of journals with incomplete data, such as incomplete annual literature or missing data in citation/Altmetrics. There were 80 journals in the study sample when the journals with deficit data were removed. The study chose the reader counts from Mendeley, literature review number from Twitter and the score number from Altmetrics.com as the representation of Altmetrics data. These three Altmetrics indices represent user numbers, influence and attention degree of article uploaded
to online reference manager, social media and Altmetrics website. The following methods were applied for collecting every specific Altmetrics data. It is convenient to apply the Webometrics Analyst 2.0 developed by the webometrics study group coming from Wolverhampton University to gain the reader counts of Mendeley (Thelwall, 2009). The Mendeley application program interface (API) was embeded in Webometrics Analyst 2.0 and Mendeley reader counts for every article can be retrieved by title, authors, DOI and something else that can identify the specific article in this data collection tool. The article score and Twitter comments was obtained by calling rAltmetrics function package in R language via article DOI. At this stage, the data obtained all belongs to paper level. The process of merging paper data to journal level need to finish. There were 12 journals without Mendeley reader counts because of no DOI in the journal articles. If there were no DOI in the article, the rAltmetrics package couldn’t be applied to collect Score data and comment data. So those journal without DOI were removed from the research. There were also another 12 journals without return results by using rAltmetrics to obtain article score and Twitter comment. This study eliminated the 12 journals in the stage of data analysis in order to ensure accuracy of analysis results. Finally, a total of 56 journals were selected as study object.

Results and Discussion

Data distribution analysis of citation and Altmetrics

Indicators (including JIF, ReaderIF, PostIF and ScoreIF) were arranged in descending order according to their numeric value and then the comparison graphs (Figure 1, 2, 3) of Journal Impact Factor and Altmetrics were shown in order to explore the distribution characteristics between citation indices and Altmetrics indices.

From the whole point of view, the three Altmetrics factors in Figure 1, 2, 3 are above JIF. The trend of descending graph in the three pictures all indicate the long tail trend. The number of journal belonged to high value division is less but a lot of journals are divided into the low value division. The indicators of the high value journal are much higher than the corresponding indices value of journals belonged to the low value region. There is no obvious difference in the evaluation indices of journals belonged low value impact factors. The Matthew effect of concentration and dispersion is more obvious on the open access platform. The high Altmetrics impact factor is much larger than the low Altmetrics impact factor in the three Altmetrics indices. Although there is a high value distribution in Journal Impact Factor, but there is no huge difference between high JIF and low JIF.

It can be seen from the figure that three Altmetrics indices are significant higher than JIF, with a large difference in the descending trend of ReaderIF and JIF. The JIF of 56 journals is lower than 5, Fig.1 shows that 17 journals’ ReaderIF are higher than 5. It is found that the biggest ReaderIF is 20.038, while the biggest JIF is 4.775, indicating the diversity revealability of journals on academic environment and network environment. The primary data was checked in order to find out which journals get the biggest ReaderIF and the biggest JIF. The journal of GOVERNMENT INFORMATION QUARTERLY gets the biggest ReaderIF, with JIF value of 2.515, which focus on policy, information technology and government research as well as the usability problems of government information. The journal of JOURNAL OF INFORMATION TECHNOLOGY gets the biggest JIF, with ReaderIF value of 4.595, which focus on management science, information system and computer science. It is found that research field of journal, literature content besides journal quality will affect the attention and visibility of journal on academic system and non-academic environment via a comparative analysis.

The differences between ReaderIF and PostIF ,ScoreIF are shown in Fig.1, Fig.2, and Fig.3. Fig.1 indicates a relatively uniform distribution for ReaderIF in high value division and shows
a stable transition from the high value to low value. Although the high ScoreIF is distributed in Fig. 3, but there is a steep transition from high value to low. Only one journal belongs to high ScoreIF. That is "JOURNAL OF COMPUTER-MEDIATED COMMUNICATION", which focus on the field of Internet, communication and network technology, multimedia and wireless communication. These research field is popular in the field of Information Science and get generous attention on social media. It is also indicates that the journal also belongs to high distribution of JIF (3.541) in the study sample. Both JIF and ScoreIF reflect the better quality of the journal. Topics and scopes of the journal may have an effect on attention and influence of the literature on online open platform. Social media, online reference manager as well as other open access platform decrease the threshold of academic access, so that more people can follow with and discuss the hot topics. Digitization and internetization for journals expand the scope of academic information exchange further more bring high Altmetrics for journals.

Figure 1: Descending comparison of ReaderIF and JIF

Figure 2: Descending comparison of PostIF and JIF

Figure 3: Descending comparison and ScoreIF and JIF
Correlation analysis of citation indices and Altmetrics indices

The normal distribution test was conducted in this study before the correlation analysis. An exploratory analysis in SPSS tool was used and the normal report is shown in Table 1. 56 journals was imported to SPSS then Shaprio-Wilk test was conducted to validate the data distribution. Table 1 shows the P value is less than 0.05, so the original data doesn’t confirm to normal distribution. Then the Spearman rank correlation analysis was applied in the next step. The coefficient between citation indices and Altmetrics indices is calculated, and the coefficient matrix is shown in figure 4.

Table 1: Tests of Normality

<table>
<thead>
<tr>
<th></th>
<th>Kolmogorov-Smirnova</th>
<th>Shapiro-Wilk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statistic</td>
<td>df</td>
</tr>
<tr>
<td>JIF</td>
<td>.144</td>
<td>56</td>
</tr>
<tr>
<td>Immediate</td>
<td>.192</td>
<td>56</td>
</tr>
<tr>
<td>Eigen</td>
<td>.274</td>
<td>56</td>
</tr>
<tr>
<td>ReaderIF</td>
<td>.211</td>
<td>56</td>
</tr>
<tr>
<td>TweeteIF</td>
<td>.206</td>
<td>56</td>
</tr>
<tr>
<td>ScoreIF</td>
<td>.231</td>
<td>56</td>
</tr>
</tbody>
</table>

Figure 4: Coefficient matrix between Citation indices and Altmetrics indices
Previously, some scholars have studied the correlation between citation and Altmetrics data and the results show that the correlation between them is low (Mohammadi, & Thelwall, 2014). Similarly, Fig. 4 displays a positive correlation between JCR citation index and Altmetrics index at the journal level, but the coefficients were less than 0.4. The relationship between ReaderIF and Immediately index indicates the highest correlation meanwhile, the relationship between ReaderIF and Eigenfactor shows the lowest correlation. However, PostIF and ScoreIF have the highest correlation with Eigenfactor, at the same time those two Altmetrics indices all show the low correlation with JIF. Difference in correlation results of citation indicators and Altmetrics indices reflects that there is a diversity between Altmetrics data, since they were collected from different network platforms though they all belong to Altmetrics.

Correlation coefficient of ReaderIF and PostIF is 0.22 and the coefficient of 0.19 between ReaderIF and ScoreIF, reflecting weak correlation at the bottom right of Fig. 4. And the correlation relationship of 0.97 between PostIF and ScoreIF, showing strong correlation. Mendeley, Twitter and Altmetrics.com are important sources of Altmetrics data, but the correlation between the three showed big differences. Studies find that Altmetrics online research Score is mainly based on three parts: The number of social media, the influence of social media, and Tweeter (Wu, S.N & Zhao, R.Y. 2016). Twitter is a popular social network and micro blogging services website with a large number of users. Twitter is one of the keynote objects when Altmetrics.com collect the article level data and occupied a considerable proportion in the article score. Therefore, a significant correlation is shown between ScoreIF and PostIF. But the Mendeley reader counts is not counted in the paper Score. In addition, as an online academic management platform, there are great differences in the aspect of functional positioning, platform property, the audience and application nature between Mendeley and social media represented by Twitter. These are the reasons for the low correlation between ReaderIF and PostIF or ScoreIF respectively. The low correlation between ReaderIF and the other two indicators suggests that Mendeley, Altmetrics.com and Twitter are social medias and they are all important data sources in Altmetrics research. However, there are diversities between different social media platforms, wrong results may happen if only one kind of Altmetrics data was used in evaluation of scientific researches. Only by taking the multi-source Altmetrics data together to evaluate the academic achievements such as papers and journals can we get more objective social media influence.

Journal impact factor, immediate index and Eigenfactor are commonly used indexes in journal evaluation, but it is shown that the coefficient of 0.61 and 0.51 between immediately index and JIF, Eigenfactor respectively in the top left hand corner of Fig. 4, reflecting moderate correlations. The coefficient of JIF and Eigenfactor is 0.88, which reflects significant correlation. Though, the three citation indices are calculated from the perspective of reference relationship between papers, there are different mechanisms between JIF, immediately index and Eigenfactor. The immediately index is calculated using the current citation data in the year of publication. However, other two citation indices is calculated using the data of journal published after two years. There is a time effect on the index based on citation relationship. Therefore, the six indicators were all introduced into the evaluation matrix in the next analysis, and the comprehensive citation index and Altmetrics index are used to evaluate the journals.

Construction journal evaluation model by using citation data and Altmetrics data

Firstly, the star graph included the three citation indices and Altmetrics indices was drawn to show the relative size of each indicator before building the model of journal evaluation. The Eigenfactor is thousand times smaller than other indicators, so the Eigenfactor can't be showed in the star graph. In order to display all the indicators in the graph, the original data was translated with the nonlinear normalization. The method of transformation is as follows (Cai, Wang, & Liu, 2015):
Data nondimensionalize processing:

\[ r_{ij} = \frac{x_{ij} - \min x_{ij}}{\max x_{ij} - \min x_{ij}} \]  (1)

Data nonlinear processing:  \[ R_{ij} = \frac{\pi}{2} \arctan(r_{ij} + 1) \]  (2)

The stars function in R was used to construct the star graph after the nonlinear processing. Here, four journals in Q1 to Q4 journal division in JCR were respectively selected as the sample representation in each journal division to display their star graphs, because it will so miscellaneous if all journal star graphs were drawn.

![Figure 5: Visualization of journal star](image)

There are six indicators in this study, the proportion of each indicator is 60 degree angle. Length of the sector indicates the relative size of each indicator. Advantage and disadvantage of every indicator are reflected by the sector area. Fig.5 reveals that the representative journal of Q1 has the largest fan-shaped area on JIF, indicating that the journal has an advantage in JIF. Similarly, the representation of Q2 shows a big occupation in the area of JIF and Reader, reflecting a good performance in these two indicators. The sample of Q3 has advantage in ReaderIF indicator compared with other indices, however, the sample of Q4 has a balance performance on each indicator.

The star graph for every journal could be displayed with the method shown in figure 5. The comprehensive performance of the six indicators can be obtained from the star graph and the overall evaluation score also can be calculated by adding up six sector areas of the star graph. Finally, the overall score is applied to assess the journals. The formula for calculating the star graph area is:

\[ S_i = \frac{1}{2} \sin 60 \left( R_{i1} R_{i2} + R_{i2} R_{i3} + \cdots + R_{i6} R_{i1} \right) \]  (3)
$R_{ij}$ represents the data nonlinear processed. The formula above was used to obtain the star area of journals. Area size was used to measure the comprehensive level for every journal from two dimensions of citation and Altmetrics. Finally, the study made the Spearman correlation analysis between the journal ranking in JCR and the ranking in this study and a high positive correlation between the two sorted results is shown in Table 2.

The journal evaluation mode in this paper not only consider the journal academic factor but also inspect the non-academic indicators of influence on social media and reader counts compared with JCR journal evaluation system. Traditional journal evaluation system doesn't contain non reference data, such as reader counts, comment number, collection data and so on, however the comprehensive journal evaluation model in this study could avoid the defectiveness that non-citation data was not included in the evaluation system. On the other hand the citation indices no longer come first in the integrated evaluation model combined with citation data and Altmetrics, because citation and non-citation data act together on the final assessment outcome. This comprehensive evaluation model, which combines academic influence factors and non-academic influence factors, can produce more objective and fair evaluation results for journals.

### Table 2: Correlations Result of journal ranking

<table>
<thead>
<tr>
<th>OrderJCR Spearman's rho</th>
<th>OrderArea</th>
<th>Correlation Coefficient</th>
<th>R</th>
<th>Sig. (2-tailed)</th>
<th>N</th>
<th>Correlation Coefficient</th>
<th>R</th>
<th>Sig. (2-tailed)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.000</td>
<td></td>
<td>.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OrderJCR</td>
<td></td>
<td></td>
<td>.827**</td>
<td></td>
<td>.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spearman's rho</td>
<td></td>
<td></td>
<td>1.000</td>
<td></td>
<td>.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OrderArea</td>
<td></td>
<td></td>
<td>.827**</td>
<td></td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.000</td>
<td></td>
<td>.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>56</td>
<td></td>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>56</td>
<td></td>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).

### Conclusion
This study analyzed the data distribution, difference and relationship between citation and Altmetrics as well as the method of applying those two dimension data to evaluate journals by using the research object of journals in Information Science & Library Science. The research reveal that both citation and Altmetrics don't confirm to normal distribution and show the long tail distribution trend. A small number of journals belong to high value division, a lot of journals are divided into the low value journal division. The three Altmetrics indicators of ReaderIF, PostIF and ScoreIF are significant higher than citation indices. At the aspect of correlation analysis, there is higher correlation with ReaderIF and immediately index, while a higher correlation is revealed between between PostIF, ScoreIF and Eigenfactor. Though the reader counts, comment numbers and article score are all Altmetrics data, but there is a big difference in the aspects of data source, platform attribute, function orientation and target audience. So many dissimilarities result in the different visibility and attention on open access platforms. In addition, the factors of academic quality, subject field and research topic also work on the attention degree of journals in network environment. There is a significant correlation between journal ranking in JCR and the ranking calculated by star area, indicating the validation of using star area to measure journals combined with citation and Altmetrics. The journal evaluation model incorporated citation data and Altmetrics could conduct the assessment from academic
perspective and non-academic perspective in order to get a relatively comprehensive and thorough outcome compared the traditional journal assessment system. This study is an exploratory analysis of using citation data and Altmetrics data via star area to evaluate journal comprehensively. There are some deficiencies in the study: Firstly, there are only Mendeley reader counts, comment numbers and article score three parts of Altmetrics as well as parts of citation data in this study, so the comprehensive performance of journals in social network and academic system can’t be fully assessed. On the other hand, only one subject of Information Science & Library Science is selected as the study object. It is also a deficiency of no analysis to journals in the other subjects.

Acknowledgements
This research is supported by the National Social Science Fund of China [Grant No. 14BTQ067].

References
Hirsch, J. E. (2007). Does the h index have predictive power?. Proceedings of the National Academy of Sciences of the United States of America, 104(49), 19193.
Nuredini, K., & Peters, I. (2016). Enriching the knowledge of altmetrics studies by exploring social media metrics for Economic and Business Studies journals. International Conference on Science and Technology Indicators.

Zahedi, Z., Costas, R., & Wouters, P. (2015). Do mendeley readership counts help to filter highly cited wos publications better than average citation impact of journals (jes)? *Computer Science.*
Alternative Metrics Correlations: Do Academic Reviews Correlate with Library Holdings?

Zhou Qingqing¹, Wang Shenghui², Zhang Chengzhi³*

¹breeze7zhou@163.com
Nanjing University of Science and Technology, Nanjing (China)
Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University), Fuzhou (China)

²shenghui.wang@oclc.org
OCLC Research, Leiden (the Netherlands)

³zhangcz@njust.edu.cn
Nanjing University of Science and Technology, Nanjing (China)
Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University), Fuzhou (China)

Abstract
Existing researches have proved that both academic reviews and library holdings can be alternative sources to assess impacts of academic books. This paper endeavors to identify correlations between the two sources, which may be useful for book ordering of libraries. Specifically, 69,263 academic reviews in Choice (Choice: Current Reviews for Academic Libraries) are collected with four metrics: recommendation levels, readership levels, numbers of interdisciplinary subjects and review contents. Then, a topic model is used to extract topics from review contents. Meanwhile, library holdings of each book are identified, including the total number of library holdings, holding regions and holding distributions based on an entropy method. Finally, correlation analysis between Choice reviews and library holdings are conducted. Experimental results reveal that books with higher recommendation levels or extensive readerships tend to be ordered more by academic libraries. Meanwhile, books with extensive readerships will be collected more uniformly by academic libraries. In conclusion, metrics derived from Choice academic book reviews can be used as indicators to recommend book ordering of academic libraries.

Keywords: book impact assessment; topic extraction; review mining; altmetric; library holdings

Conference Topic
Altmetrics

* Corresponding author, Email: zhangcz@njust.edu.cn
Introduction

Traditional book impact assessments are based on citations, such as Krampen, Becker, Wahner, and Montada (2007), Barilan (2010) etc. With the rapid development of Web 2.0, more alternative metrics are used to assess qualities of books. Donovan and Butler (2007) proved that publisher prestige can be proposed as an alternative method for the impact assessment of books. According to Torres-Salinas and Moed (2009) and Zuccala, Verleysen, Cornacchia, and Engels (2014), library holdings can give useful information for evaluating qualities of books. Shema, Bar-Ilan, and Thelwall (2014) used blog citations to measure impacts of books. In a bid to obtain more comprehensive assessment results, researchers endeavored to mine content information about books, such as Kousha and Thelwall (2016), and Zhou, Zhang, Zhao, and Chen (2016).

On the basis of Kousha and Thelwall (2015) and White et al. (2009), academic reviews and library holdings are useful sources for measuring book impacts. In this paper, we try to identify correlations between these two sources via a fine-grained analysis, which may be conducive for book ordering of libraries. Specifically, academic reviews in Choice (Choice: Current Reviews for Academic Libraries) are collected. Each review contains four metrics: recommendation level, readership level, number of interdisciplinary subject and review content. Then, we extract topics from review contents via a topic model. Meanwhile, we collect the global and regional library holding information of each book, and analyze their distributions using an entropy method. Finally, the correlation analysis between Choice reviews and library holdings shows that books with higher recommendation levels or extensive readerships will be ordered more by academic libraries. In addition, books with extensive readerships tend to be collected more uniformly by academic libraries, while books with higher recommendation levels or more topics in reviews may be non-uniformly distributed among different libraries. Hence, we conclude that metrics derived from Choice academic reviews can give suggestions to recommend book ordering to academic libraries.

Related works

This paper tries to identify correlations between two alternative metrics of book impact assessment: academic reviews and library holdings, and analyzes review contents with a topic model. Hence, in this section, we describe two categories of related works: academic book impact assessment and topic extraction.

Academic book impact assessments

Citation is a valuable measurement for evaluating impacts of academic books. Kousha, Thelwall, and Rezaie (2011) proved online citations from Google Books and Google Scholar can be sufficiently numerous to support peer review for research evaluation, according to significant correlations (about 0.7) among the three metrics: Google Books, Google Scholar and Scopus. Torressalinas, Robinsongarcia, and Lopezcozar (2012) analyzed different impact indicators referred to scientific books in Social Sciences and Humanities fields during 2006-2011 via Book Citation Index. Gorraiz,
Purnell, and Glänzel (2013) indicated that the BKCI was a first step toward creating a reliable and necessary citation data source for book impact assessments. Abrizah and Mike (2014) presented that Google Books and Google Scholar can be recommended for impact assessments of non-Western books, as 23% of 1357 books in Google Books, 37% in Google Scholar had a higher proportion cited if they were older or in English. Citation-based methods are becoming inadequate for Web 2.0, thus researchers are seeking more online information to assess book impacts. White et al. (2009) used libcitation count to assess book impact. CabezasClavijo et al. (2013) took library loans statistics as a measure of book impact. Shema et al. (2014) proved that blog citations can be used as an alternative metric for book impact assessments, according to statistically significant evidence (about 0.04 correlations between the blogged metrics and non-blogged metrics) for articles published in 2009 and 2010.

In order to assess impacts of academic books more comprehensively, content information are analyzed. Kousha and Thelwall (2016) proved that metrics derived from Amazon.com reviews of academic books could give evidence about their impacts, according to low but significant correlations (about 0.2) between amazon metrics and other metrics. Zhou et al. (2016) indicated that Amazon.cn reviews can be used to assess academic book impacts based on the significant correlations (about 0.4) between citations versus Amazon reviews.

From analysis above we can conclude that researchers are mining more interesting and comprehensive methods to obtain impacts of academic books.

**Topic extraction**

Topic modeling algorithm is a powerful and computational tool to help organize and summarize information at a scale that would be impossible by human (D. Blei, Carin, & Dunson, 2011). Classic topic modeling algorithms include latent semantic analysis/indexing (LSI) (Deerwester, 1990), probabilistic latent analysis/indexing (PLSI) (Hofmann, 1999), latent dirichlet allocation (LDA) (D. M. Blei, Ng, & Jordan, 2003) and hierarchcal dirichlet processing (HDP) (Wang, Paisley, & Blei, 2011). Advantages of each algorithm are unique. Specifically, LSI algorithm can depict the synonyms effectively, while PLSI algorithm is based on LSI algorithm and quite capable of distinguishing polysemy. Benefit from Bayesian Framework, LDA algorithm reduces over fitting significantly. Different from three algorithms above, HDP algorithm can identify topic numbers automatically.

As an effective tool, topic models are applied in many ways. Some researches focus on extracting topics in social media. Cataldi, Caro, and Schifanella (2010) proposed a topic detection technique to detect emerging topics on Twitter based on temporal and social terms evaluation, which permitted to retrieve in real-time the most emergent topics expressed by the community. Madani, Boussaid, and Zegour (2015) also focused on identifying topics expressed by users. Topic models can also be used to conduct users’ behavior survey. Roberts et al. (2014) used the structural topic model to make analyzing open-ended responses easier and more revealing. Huynh, Fritz, Mario, Schiele, and Bernt (2015) discovered activity patterns in a user's daily routine with topic models automatically.
In the present study, we aim to find correlations between academic reviews and library holdings. Differs from Kousha and Thelwall (2015), which used number-based metrics in Choice only, review contents are mined in this paper. Hence, we may get more details from Choice reviews. In contrast to existing content-level researches, which analyzed online book reviews with sentiment classification, we focus on extracting topics expressed by reviewers. It may be more suitable for analyzing contents with less sentiment expressions, such Choice academic reviews.

Research Questions
The following research questions are designed to assess whether metrics from *Choice* academic reviews are correlated to library holding statistics.
1. Are there correlations between Choice metrics, including recommendation levels, readership levels, numbers of interdisciplinary subjects and review contents?
2. Do Choice metrics correlate with library holding indicators, including library holding numbers, regions and distributions?

Method

*Choice academic reviews*
69,263 book review information from Choice Reviews Online were extracted from the *Humanities*, *Reference*, *Social & Behavioral Sciences*, and *Science & Technology* categories. Each review includes information about *title*, *author*, *year*, *review content*, *reviewer*, *recommendation level*, *readership level*, *interdisciplinary subjects* and *subject* (Table 1).

<table>
<thead>
<tr>
<th>Table 1. An example of Choice book reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Author</strong></td>
</tr>
<tr>
<td><strong>Year</strong></td>
</tr>
<tr>
<td><strong>Review content</strong></td>
</tr>
<tr>
<td><strong>Reviewer</strong></td>
</tr>
<tr>
<td><strong>Recommendation level</strong></td>
</tr>
<tr>
<td><strong>Readership Level</strong></td>
</tr>
<tr>
<td><strong>Interdisciplinary Subjects</strong></td>
</tr>
<tr>
<td><strong>Subject</strong></td>
</tr>
</tbody>
</table>

*Choice recommendation levels*
The recommendation levels assigned to Choice reviews were converted into a number,
from 1 for ‘Not recommended’ to 5 for ‘Essential’ (Table 2).

<table>
<thead>
<tr>
<th>Recommendation levels</th>
<th>Not recommended</th>
<th>Optional</th>
<th>Recommended</th>
<th>Highly Recommended</th>
<th>Essential</th>
</tr>
</thead>
<tbody>
<tr>
<td>numbers</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Readership Level

The readership levels assigned to Choice reviews were converted into a number, from 1 for ‘All Readership Levels’ to 8 for ‘Professionals/Practitioners’ (Table 3). As a book may have different readership levels, we make the lowest one as the final readership level, some examples are shown in Table 4.

<table>
<thead>
<tr>
<th>Readership Level</th>
<th>numbers</th>
<th>Readership Level</th>
<th>numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Readership Levels</td>
<td>1</td>
<td>Graduate Students</td>
<td>5</td>
</tr>
<tr>
<td>General Readers</td>
<td>2</td>
<td>Researchers/Faculty</td>
<td>6</td>
</tr>
<tr>
<td>Lower-division Undergraduates</td>
<td>3</td>
<td>Two-Year Technical Program</td>
<td>7</td>
</tr>
<tr>
<td>Upper-division Undergraduates</td>
<td>4</td>
<td>Professionals/Practitioners</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 4. Examples of readership levels in Choice book reviews

<table>
<thead>
<tr>
<th>Readership Levels</th>
<th>numbers</th>
<th>Final number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graduate Students, Researchers/Faculty, Professionals/Practitioners</td>
<td>5, 6, 8</td>
<td>5</td>
</tr>
<tr>
<td>All Readership Levels, General Readers, Lower-division Undergraduates, Graduate Students, Researchers/Faculty, Two-Year Technical Program Students, Professionals/Practitioners</td>
<td>1, 2, 3, 4, 5, 6, 7, 8</td>
<td>1</td>
</tr>
<tr>
<td>Upper-division Undergraduates, Graduate Students</td>
<td>4, 5</td>
<td>4</td>
</tr>
</tbody>
</table>

Interdisciplinary subjects

The numbers of interdisciplinary subjects were counted for each book, as shown in Table 5.

<table>
<thead>
<tr>
<th>Interdisciplinary subjects</th>
<th>numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian and Asian American Studies</td>
<td>1</td>
</tr>
<tr>
<td>Social Behavioral Sciences - Education</td>
<td>2</td>
</tr>
<tr>
<td>Humanities - Religion</td>
<td>2</td>
</tr>
<tr>
<td>Science Technology - Biology - Botany</td>
<td>3</td>
</tr>
</tbody>
</table>

Review contents

A topic model was used to analyze topics expressed in review contents. In this paper, we used Latent Dirichlet Allocation (LDA) (Hoffman, Blei, & Bach, 2010) to identify topics. As one review may express several topics, a threshold was used to filter topics with lower probabilities. We compared different topic numbers and thresholds, and
got optimal performances when the number was 20 and the threshold equaled 0.1. Table 6 represents samples of topics in Choice book reviews.

**Table 6. Examples of topics in Choice book reviews**

<table>
<thead>
<tr>
<th>Review contents</th>
<th>Topics</th>
<th>Topic numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>This volume tells the story of one woman’s development as a teacher and offers a very insightful look at the process of learning to teach. It is based on a significant amount of information about the book's central figure, Kay: her graduate-level teacher preparation classes through her own memories, papers she wrote while studying to become a teacher, her student teaching journal, observations of her in action with children and supervising teachers, and important conversations … This volume does an admirable job of reviewing the central question in understanding how we can best prepare, support, and sustain deliberative teachers.</td>
<td>1. women</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2. occupation</td>
<td></td>
</tr>
</tbody>
</table>

In 1445, John II of Castile granted a powerful noble supporter some land and related jurisdiction that the city of Toledo had purchased earlier… He emphasizes the centrality of judicial institutions and due process in "good government" by analyzing in detail the documentation required by judicial procedures. Owens argues that Castilians who observed their monarch adhering to due process and principles of justice formed a cornerstone of support for royal authority. Monarchs who employed "absolute royal authority" in opposition to generally held legal principles, however, undermined their position…

<table>
<thead>
<tr>
<th>ISBN</th>
<th>Library holding information</th>
</tr>
</thead>
<tbody>
<tr>
<td>9810222***</td>
<td>‘2 @ Denmark’, ‘1 @ France’, ‘2 @ Germany’, ‘1 @ Netherlands’, ‘2 @ Sweden’, ‘2 @ Switzerland’, ‘3 @ United States of America’</td>
</tr>
<tr>
<td>9789971697***</td>
<td>‘1 @ New Zealand’, ‘3 @ United States of America’</td>
</tr>
<tr>
<td>9789971696***</td>
<td>‘1 @ China’, ‘2 @ Denmark’, ‘1 @ France’, ‘2 @ United States of America’</td>
</tr>
</tbody>
</table>

*Library holdings*

Library holding information of 61933 books from WorldCat.org (OCLC) were extracted (7330 of 69263 books have no holding information in OCLC), including library holding numbers of books in each region (see Table 7).

**Table 7. Library holding information from OCLC**

<table>
<thead>
<tr>
<th>ISBN</th>
<th>Library holding information</th>
</tr>
</thead>
<tbody>
<tr>
<td>9810222***</td>
<td>‘2 @ Denmark’, ‘1 @ France’, ‘2 @ Germany’, ‘1 @ Netherlands’, ‘2 @ Sweden’, ‘2 @ Switzerland’, ‘3 @ United States of America’</td>
</tr>
<tr>
<td>9789971697***</td>
<td>‘1 @ New Zealand’, ‘3 @ United States of America’</td>
</tr>
<tr>
<td>9789971696***</td>
<td>‘1 @ China’, ‘2 @ Denmark’, ‘1 @ France’, ‘2 @ United States of America’</td>
</tr>
</tbody>
</table>

In order to analyze library holding distributions of books in different regions, an entropy method was used (Hongzhan, Lü Pan, & Yao, 2009), which can be computed by Eq. (1) - (2). If library holdings of a book are uniformly distributed, the entropy value of the book will be higher.
\[ e = -\frac{1}{\ln(N)} \sum_{j=1}^{N} p_j \ln(p_j) \]  

(1)

\[ p_j = \frac{v_j}{\sum_{j=1}^{N} v_j} \]  

(2)

Where, \( v_j \) is the holding numbers of the book in region \( j \), and \( N \) is the number of regions. Then, we obtained distribution information of books in Table 7 (see Table 8).

**Table 8. Library holding statistics from OCLC**

<table>
<thead>
<tr>
<th>ISBN</th>
<th>Numbers</th>
<th>Regions</th>
<th>Distributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>9810222***</td>
<td>13</td>
<td>7</td>
<td>0.967</td>
</tr>
<tr>
<td>9789971697***</td>
<td>4</td>
<td>2</td>
<td>0.811</td>
</tr>
<tr>
<td>9789971696***</td>
<td>6</td>
<td>4</td>
<td>0.959</td>
</tr>
</tbody>
</table>

**Results**

**Correlations analysis between Choice metrics**

There are significant Spearman correlations among the three metrics, recommendation levels, readership levels and numbers of interdisciplinary subjects (Table 9). Specifically, recommendation levels have negative Spearman correlations (0.261) with readership levels. Hence, books with lower readership levels tended to be recommended more highly by book reviewers. Meanwhile, correlations between readership levels and interdisciplinary subjects are low but significant. Hereby, we can conclude that interdisciplinary books may be suitable for readers of higher levels.

**Table 9. Spearman correlations between Choice metrics (n=69263).**

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Recommendation levels</th>
<th>Readership levels</th>
<th>Interdisciplinary subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommendation levels</td>
<td>1</td>
<td>-0.261**</td>
<td>-0.023**</td>
</tr>
<tr>
<td>Readership levels</td>
<td></td>
<td>-0.261**</td>
<td></td>
</tr>
<tr>
<td>Interdisciplinary subjects</td>
<td>1</td>
<td>0.037**</td>
<td></td>
</tr>
</tbody>
</table>

**. Significant at p=0.05

**. Significant at p=0.01

**Figure 1. Distribution of topic numbers (n=69263).**

In order to mine more details about Choice academic reviews, we use topic models to identify topics expressed in the review contents via LDA. Then, we get numbers of
topic in each review. The descriptive statistics information about topic numbers is represented in Figure 1. The horizontal axis of Figure 1 represents topic numbers in reviews and the vertical axis represents the number of reviews. Figure 1 shows that the number of topics in individual review ranges from 1 to 5, and most reviews contain 3 topics.

Then, we analyzed correlations between topic numbers and Choice metrics (Table 10). It can be seen from Table 10 that, the number of topics correlates positively with recommendation levels but negatively with readership levels and numbers of interdisciplinary subjects. This suggests that books with more topic numbers in reviews may be recommended more highly by book reviewers. Meanwhile, reviews tend to express less topics when comment interdisciplinary books. In addition, if a book can be read by lower level readers, reviewers may express more topics in the reviews. In addition, we can see correlations in Table 10 are significant but low, which indicate that although there are correlations between topics and other three Choice metrics, they cannot replace topics. Therefore, we can conclude that it is necessary to mine review contents.

Table 10. Spearman correlations between topic numbers and other three Choice metrics (n=69263).

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Recommendation levels</th>
<th>Readership levels</th>
<th>interdisciplinary subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic numbers</td>
<td>0.009**</td>
<td>-0.012**</td>
<td>-0.040**</td>
</tr>
</tbody>
</table>

Correlations analysis between Choice metrics and library holding indicators

The descriptive statistics information about holding indicators is shown in Figure 2. The horizontal axis of Figure 2(left) represents holding numbers and the vertical axis represents the number of books. From Figure 2(left) we can conclude that most books have less than 500 holdings, only few books have more than 2000 holdings. Figure 2(right) represents holding regions of books. From Figure 2(right) we can see that most books have been collected by less than 20 regions, few books have been collected by more than 60 regions.

Figure 2. Descriptive statistics information about holding numbers and regions (n=61933).

Figure 3 shows descriptive statistics information about holding distributions. The horizontal axis of Figure 3 represents holding distributions of books (namely, entropy
values of books, 1 means books are uniformly distributed among all regions), and the vertical axis represents book numbers. From Figure 3 we can conclude that quite a few books are uniformly distributed, and most entropy values of books are lower than 0.5.

![Distributions](image)

**Figure 3.** Descriptive statistics information about holding distributions (n=61933)

**Table 11.** Spearman correlations between Choice metrics and holding indicators (n=61933).

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Recommendation levels</th>
<th>Readership levels</th>
<th>Interdisciplinary subjects</th>
<th>Topic numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numbers</td>
<td>0.193**</td>
<td>-0.199**</td>
<td>0.014**</td>
<td>-0.002</td>
</tr>
<tr>
<td>Regions</td>
<td>0.075**</td>
<td>0.092**</td>
<td>0.020**</td>
<td>-0.022**</td>
</tr>
<tr>
<td>Distributions</td>
<td>-0.120**</td>
<td>0.310**</td>
<td>0.031**</td>
<td>-0.028**</td>
</tr>
</tbody>
</table>

There are low but significant Spearman correlations between Choice metrics and library holding indicators (Table 11). Specifically, recommendation levels and have positive correlations with holding numbers (0.193), while readership levels have negative correlations with holding numbers (-0.199). It suggests that books with higher recommendation levels tend to be ordered more by academic libraries. Meanwhile, if the readership level of a book is lower, it will get more library holdings. In other word, books with extensive readerships will be collected more by academic libraries. In addition, we can see Choice review contents have no significant correlation with holding numbers. We identified correlations between the two metrics with Figure 4. The horizontal axis of Figure 4 represents topic numbers in book reviews and the vertical axis denotes holding numbers. We can find from Figure 4 that books with 3 topic numbers in reviews tend to have higher holding numbers.

Regarding holding regions, all three Choice metrics have positive Spearman correlations with holding regions, except topic numbers.

For holding distributions, all Choice metrics have significant correlations with holding distributions. Readership levels have positive correlations with holding distributions (0.310), while recommendation levels have negative correlations with holding distributions (-0.120). It can be conclude that books with extensive readerships tend to be collected more uniformly by academic libraries. Meanwhile, books with higher recommendation levels may be non-uniformly distributed.
Discussion

This paper analyzed correlations between academic reviews and library holding information. Compared with number-based analysis methods, our approach makes use of content information through the application of topic model. It is thus not simply an analysis of correlations between numerical metrics, but also an attempt to discover reviewers’ intentions and concerns. In addition, our method is not limited to the analysis of academic books, and can be applied to other forms of books, as well as to other academic entities, such as journals, articles, research papers, reports, and so on.

According to the correlation results among Choice metrics (Table 9 and Table 10), we can find the correlations are significant but low, which ranges from 0.009 to 0.261. Therefore, we can get some conclusions like books with lower readership levels tended to be recommended more highly by book reviewers, but more important is that all the four metrics are unique and irreplaceable, as they expressed opinions of reviewers from different aspects.

Regarding correlations between Choice metrics and holding indicators (Table 11), we also get low but significant correlations, which ranges from 0.014 to 0.310. Some suggestions can be provided for libraries, such as order more books with higher recommendation levels or lower readership levels. However, it is worth noting that libraries may take above correlations as references for ordering books, but in fact, ordering decisions of libraries are based on information from more sources, which may be more objective and comprehensive compared to Choice review based only, as the Choice reviews are from one single source and subjectively generated by reviewers. Hence, metrics derived from Choice reviews may give some information for book ordering, while they are not determinants.

Our study is subject to a few limitations. First, we analyzed the relations across all fields without considering disciplinary differences, which may ignore some detailed information. We collected Choice reviews of book from four disciplines, which often have different characteristics and may result to different correlations. Thus, disciplinary differences should be studied in the future. Secondly, academic reviews came from Choice only and thus may lack the diversity of reviews available from other sources. How best to collect and integrate academic reviews from different
sources is a challenging question for future research. Thirdly, the technologies of topic model extraction need to be improved. In this study, we used a classical algorithm to extract topics expressed by reviewers. However, the rapid development of technologies for natural language processing, some advanced topic model algorithms are presented. Hence, more algorithms will be tested in the future too.

**Conclusion**

This study is to our knowledge the first to analyze correlations between academic reviews and library holding information at content level. The analysis of 69,263 books found weak but often significant relationships, suggesting that Choice academic review could be a helpful reference for libraries to ordering books.

In answer to the first research question, reviewers may express more topics in the reviews when commenting books suitable for lower level readers or non-interdisciplinary books, and they often tend to recommend these books highly. In answer to the second research question, books with higher recommendation levels or extensive readerships will be ordered more by academic libraries. Meanwhile, interdisciplinary books or books with extensive readerships tend to be collected more uniformly by academic libraries, while books with higher recommendation levels or more topics in reviews may have non-uniform distributions in different libraries.

In conclusion, our method may give some suggestions to libraries on making ordering decisions, but more evidence is needed before they could be used as tools for peer judgments about individual books. The theoretical implication of our study is that future analysis of informetrics should incorporate content within the online information. In practice, our method leverages the online information and traditional library information of books, which helps libraries to make better decisions in terms of book ordering.

**Acknowledge**

This work is supported by Major Projects of National Social Science Fund (No. 16ZAD224), Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (No. MJUKF201704) and Qing Lan Project.

**References**


statistics as a proxy for monograph selection in citation indexes. *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference* (pp. 1237-1252)


an Analysis of Scientific Co-author Network of Virtual Technology in China

Wen Ting Xiao¹  Yang Zhong²  Li Xiang³

¹593710595@qq.com
Central South University (China)

²yz_gk630@163.com
Xiangtan University ,Xiangtan (China)

³18873219089@163.com
Xiangtan University ,Xiangtan (China)

Abstract
In this paper, we collected the papers of virtualization research in the VIP database (CSCD source journals) from 1989 to 2016. Co-authorship network of China's virtual technology research was analyzed by analysis of basic measurement, community, small world effect and Scale-free phenomena by means of the social network analysis method. We found: 1. The Scale of Co-authorship network is little, the network is sparse and the component is more and the module is high, there is more cooperation within the community. 2. The co-authorship network expresses the characteristic of Small World effect, but scale-free property is not obvious, and we give the experimental results.

Conference Topic
Social network analysis

Introduction
Cooperation is ubiquitous in human society. It is an interactive process where people coordinate with each other in action for shared interests or their own benefits (Tan, 2000). In the era of big science, cooperation in scientific research is increasingly becoming an important factor influencing scientific production capacity, and cooperative research has become a universal phenomenon of modern scientific research (Huang, 2015). Through the research on the number of Co-author papers between China and other countries, Tianwei He found that the cooperative articles show exponential growth in the volume (He, 2009). Narin (1991) and Bárbara’s (2012) study found that co-author papers get more references than single author papers, and scientific research cooperation has become a common phenomenon and the underlying trends in modern scientific research.

Social network analysis, which originated from psychology, anthropology and sociology and other disciplines, is a kind of interdisciplinary research method (Knoke & Yang, 2015). In recent decades, social network is increasingly favored by the public and academia. In 1967, the "six degrees of separation" hypothesis was confirmed by the experiment of Stanley Milgrom (Watts, 1999). At the moment, social network analysis has been widely applied in a complex social system, such as mathematics, physics and all kinds of natural phenomena. With the emergence and development of UCient, Gephi, Pajek, VOSViewer, which are the social network visualization analysis tools, social network analysis method wins popularity among scholars. It is found that related research achievement about social network analysis increases rapidly by WoS retrieval. And SNA has developed into an institutionalized interdisciplinary research perspective (Knoke & Yang, 2015).

Scientific research paper is an important symbol to reflect the scientific research achievements of scientific research personnel and scientific research team. The research on author attribute in scientific research paper is the effective way to learn constitutions of
scientific research personnel, development trends and research frontier in different industries. Scientific research cooperation network was created by the scientific research cooperation, so co-author is a reflection of scientific research cooperation network. Scientific research co-author network as one of the most important social network has been widely used in determining the structure of scientific cooperation and status of individual researchers. Sarıgöl E who once used social network analysis method to study the co-author phenomenon in computer sector has found that author’s status has closely related to the success in terms of reference in the future in the scientific cooperation network (Sarıgöl, Pfitzner, Scholtes, Garas & Schweitzer, 2014); Sun Jianjun, Liu Chengbo and others analyzed scientific cooperation phenomenon on their focus field by using the method of SAN and revealed many valuable rules (Huang, 2015). The social relation of co-author network is stronger than that of citation network. References are made when the author is unknown, and can be extended across time and space; however, co-author works once formed, its time and community relations have been fixed, so it is more suitable for social network analysis (Liu, Bollen, Nelson & Sompel, 2005).

With the development of cloud computing and big data technology, virtualization technology has emerged and developed quickly (Lin, 2014). The essence of virtualization technology is to separate software application from bottom hardware and turns physical resources into logical and manageable resources. Therefore, virtualization technology is one of the most core technologies of cloud computing and big data, gaining wide attention in the fields of cloud computing and big data. To reflect comprehensively the co-author network and research status in the field of virtualization technology in China, this paper aims to measure the structure, characteristics and changes of virtualization technology research co-author paper network by using social network analysis method, providing a reference for further research of virtualization technology in China.

**Methods**

The data samples of this study is derived from VIP Database, with keywords "virtualization" as the theme for retrieval, retrieval time slot from 1989 to 2016. In order to focus on the virtualization research in the fields of cloud computing technology, the scope of journal was limited to “the CSCD source journal “, and data retrieval time slot is on January 20, 2017. After data filtering and deduplication, a total of 709 research papers were received.

Research hypothesis: We assume that the names of the authors of the paper are different in fields of virtualization technology research, and it is not to differentiate the weight according to the signature sequence of the authors. That is to say, the weight of every author in one article is the same.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Co-author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper1</td>
<td>author V1, author V2, author V3</td>
</tr>
<tr>
<td>Paper2</td>
<td>author V4, author V5</td>
</tr>
<tr>
<td>Paper3</td>
<td>author V1, author V5</td>
</tr>
<tr>
<td>Paper4</td>
<td>author V5, author V4</td>
</tr>
</tbody>
</table>

![Figure 1. samples of co-author paper network](image.png)

Network construction: node of co-author network represents the author; the line in the network represents the cooperation relationship among the authors; If there is cooperative relations between the two, then add a line between two people; If there is no cooperative
relationship between the two, then add no line between two people; Line value represents the co-author times or intensity between the authors, as shown in figure 1.

In this paper, we use Gephi and R for data analysis and visual display. Gephi is an open source free cross-platform complex network analysis software on the basis of JVM, which can do exploratory data analysis, link analysis, social network analysis and biological network analysis in a variety of networks and complex system. R is a kind of free software environment for statistical computing and graphics, and is now one of the most popular platforms for data analysis and visualization. Internet has lots of free open source R package that be competent for any complex data analysis. Original data is preprocessed by Python self-compiling program in this article. First, we extracted the author of each article; Then extracted the node and edge and assigned weight to the edge in accordance with times of cooperation, forming undirected weighted co-author network; Finally, it can become recognizable data files to Gephi and R for further analysis and processing.

Results, part I: Cooperation Network

Co-author network structure and co-author number distribution

China’s co-author network of virtualization technology research is shown in figure 2, and nodes of identical color represent the same community; line color means the earliest time that relationship was established. The distribution of the number of co-authored papers is shown in figure 3, which includes 57 single author papers, 652 co-authored papers. Coauthor rate is \((1-57) / 709 = 91.96\%\); Papers that have co-author of 2-4 accounts for the most part of the total number of papers; There are the largest number of 10 cooperators who are in a co-author paper.

Analysis of the static characteristics of virtualization technology research co-author network will help us to know cooperative situation of related subject in this field; Although figure 1
clearly presented the whole situation of the cooperation with authors in virtualization technology in China, it is difficult to clearly identify the dynamic evolution process of cooperation network. In addition, network characteristics cannot be analyzed by figure 1 in different periods to reveal its law of development. Foreign scholars such as A.L Barabási, Tomassini M, domestic scholars such as Wen Wanting, Li Chengqing, they discuss the dynamic evolution and the law of development of cooperative network on the basis of the static network analysis, and they think that the essence of the analysis is to regard the dynamic evolution network as a general social network by measuring and analyzing the static characteristics and the topology structure of the network in different periods, so as to explore the change law and characteristics of network structure (Feng & Zhao, 2014).

Table 1 shows that there are little correlative papers of virtualization technology that published in CSCD source journal before 2008. In 1998, the first paper of this subject area was published by Zhu Zhigang, Xu Guangci, Shi Dingji of Tsinghua University. And the research topic is virtual reality, then related research papers is on the increase year by year, which are in a downward trend in recent years. The structure and characteristics of virtualization technology research co-author network can be reflected by the main indexes of the network. China’s whole situation of virtualization technology research co-work network and all the main indexes data of co-author network in 1998-2016 are shown in table 1.

As is known to table 1 and figure 2, the co-author network is not fully connected. The density is 0.002, node tie is sparse and overall relationship is loose. There are 1780 authors involved in scientific creation of virtualization technology area. Although some groups are closely related in the network, there are fewer opportunities for the author groups to cooperate. This may be associated with the wide aspects and fields of virtualization technology research, such as network virtualization, storage virtualization, virtual machine and virtual reality, etc.

Table 1 Main indexes of virtualization technology research co-author network

<table>
<thead>
<tr>
<th>Year</th>
<th>Paper number</th>
<th>Nodes (N)</th>
<th>edges (E)</th>
<th>density</th>
<th>average degree</th>
<th>average path length</th>
<th>average clustering coefficient</th>
<th>Modularity</th>
<th>componenent number</th>
<th>componenent number / nodes</th>
<th>Nodes of largest component</th>
<th>isolated point number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1.0000</td>
<td>1.00</td>
<td>1</td>
<td>1.00</td>
<td>-</td>
<td>1</td>
<td>33.33%</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1999</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1.0000</td>
<td>1.00</td>
<td>1</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>50.00%</td>
<td>2</td>
</tr>
<tr>
<td>2000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2001</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0.3330</td>
<td>0.667</td>
<td>1</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>66.67%</td>
<td>2</td>
</tr>
<tr>
<td>2002</td>
<td>2</td>
<td>13</td>
<td>10</td>
<td>0.4000</td>
<td>1.600</td>
<td>1</td>
<td>1.00</td>
<td>0.600</td>
<td>0.375</td>
<td>2</td>
<td>15.38%</td>
<td>4</td>
</tr>
<tr>
<td>2003</td>
<td>10</td>
<td>22</td>
<td>30</td>
<td>0.1169</td>
<td>2.545</td>
<td>2</td>
<td>1.48</td>
<td>0.788</td>
<td>0.726</td>
<td>7</td>
<td>31.82%</td>
<td>5</td>
</tr>
<tr>
<td>2004</td>
<td>13</td>
<td>37</td>
<td>55</td>
<td>0.0796</td>
<td>2.865</td>
<td>2</td>
<td>1.89</td>
<td>0.867</td>
<td>0.811</td>
<td>9</td>
<td>24.32%</td>
<td>8</td>
</tr>
<tr>
<td>2005</td>
<td>13</td>
<td>37</td>
<td>48</td>
<td>0.0706</td>
<td>2.595</td>
<td>3</td>
<td>2.07</td>
<td>0.846</td>
<td>0.834</td>
<td>10</td>
<td>27.03%</td>
<td>8</td>
</tr>
<tr>
<td>2006</td>
<td>19</td>
<td>51</td>
<td>62</td>
<td>0.0455</td>
<td>2.275</td>
<td>2</td>
<td>1.73</td>
<td>0.783</td>
<td>0.839</td>
<td>16</td>
<td>31.37%</td>
<td>7</td>
</tr>
<tr>
<td>2007</td>
<td>18</td>
<td>51</td>
<td>74</td>
<td>0.0525</td>
<td>2.627</td>
<td>2</td>
<td>1.57</td>
<td>0.756</td>
<td>0.818</td>
<td>15</td>
<td>29.41%</td>
<td>6</td>
</tr>
<tr>
<td>2008</td>
<td>27</td>
<td>73</td>
<td>97</td>
<td>0.035</td>
<td>2.521</td>
<td>2</td>
<td>1.56</td>
<td>0.808</td>
<td>0.891</td>
<td>21</td>
<td>28.77%</td>
<td>5</td>
</tr>
<tr>
<td>2009</td>
<td>30</td>
<td>80</td>
<td>119</td>
<td>0.0342</td>
<td>2.75</td>
<td>2</td>
<td>1.7</td>
<td>0.735</td>
<td>0.849</td>
<td>24</td>
<td>30.00%</td>
<td>7</td>
</tr>
<tr>
<td>2010</td>
<td>45</td>
<td>152</td>
<td>249</td>
<td>0.0207</td>
<td>3.145</td>
<td>2</td>
<td>1.71</td>
<td>0.914</td>
<td>0.92</td>
<td>40</td>
<td>26.32%</td>
<td>10</td>
</tr>
<tr>
<td>2011</td>
<td>85</td>
<td>258</td>
<td>418</td>
<td>0.019</td>
<td>3.116</td>
<td>3</td>
<td>2.09</td>
<td>0.847</td>
<td>0.925</td>
<td>68</td>
<td>26.36%</td>
<td>12</td>
</tr>
</tbody>
</table>
Basic indexes include the number of papers, network scale (N), the number of network edges (E), network density and average degree. The number of nodes and edges of virtualization technology research in China is on the rise after 2002 and is on the downward trend after 2012. 2011~2012 is the peak period of cloud computing research in China. After that, the market pays less attention on virtualization and relevant technology has stepped into maturity, the focus of research is beginning to shift. Compared with network density, average degree is not restricted by the scale of network and has an advantage in terms of coherence measure, so the average degree is often used for comparison among networks with different node number (Nooy, Mrvar & Batagelj, 2012). From 1998 to 2002, our average degree of virtualization technology co-author network is about 1. From 2002 to 2009, the average degree is at around 2.5; the average degree from 2010 to 2016 is at around 3.1. It is can be seen that virtual technology research co-work network did not fluctuate much between 1998 and 2016.

Network centrality index is composed of diameter, average path length and average clustering coefficient. Diameter refers to the maximum geodesic distance in all node pairs (the shortest path between two nodes). In scientific cooperation network, network diameter represents the maximum distance that can reflect the relationship establishment between researchers. Average path length (L) refers to average value among all node pairs, namely $L = \frac{2}{N(N-1)} \sum_{i \neq j} d(i,j)$, $d(i,j)$ represents the shortest path between nodes, which can characterize the transmission performance and efficiency of the network. Average clustering coefficient(C) represents average probability that two nodes connecting to the same node are interconnected, calculation formula is $C = \frac{1}{N} \sum_{i=1}^{N} \frac{2E_i}{k_i(k_i-1)}$. Average clustering coefficient depicts the transitivity of the network, which describes the concentration of researchers and reflects the tightness of co-author relationship. From table 1, the network diameter is less than or equal to 3, and average path length are smaller than 2.1 between 1998 and 2016, which shows that there is a close relationship between researchers in the subject field, and information and knowledge can be flown and transfer effectively in the group; Average clustering coefficient is very high and there is no large fluctuation on the whole. In other words, it has very compact and relatively stable co-author groups and there is smooth flow of knowledge in groups.

Connectivity index is composed of component number, the node number proportion in constituent element, the node number in the largest component and isolated point number. Modularity can evaluate the quality of component, and there is a close link between components of high modularity in vertex of the same group, but there is only sparse connection of vertex in different groups. The distribution of component node number in virtualization technology research co-author network in China is shown in figure 4, and about 90% of the component node number is less than 10, the former three component node number is 71, 57, 54 respectively; From the observation of "node number of the largest component" from connectivity index, the biggest component of co-author network in each year has only 12 members. We can infer that small component each year can gradually merge into a larger component as the variation with time, and it represents that the scope and depth of scientific cooperation continue to expand. Although there is a lot of co-author network component and high modularity, the bigger community scale is relatively small. Small research teams in the field of virtualization technology are ubiquitous in the virtualization technology research field;
there is a close relationship in internal cooperation of small groups. However, there is loose relation of each group, and this is the view of virtualization technology having many aspects to study.

![Figure 4 the distribution of component node number](image)

**Results, part II: community**

"Birds of a feather flock together", community structure is a basic feature of each kinds of network in real life. Its basic performance is that there is dense edge between node and node within the communities and sparse between the communities. In scientific cooperation network, there is more cooperation among scientific research personnel within the communities and less cross-community cooperation. The topology structure and the overall function of the network is the result from joint effects in each community, which detects community structure in virtualization technology research fields and helps to identify team belonging of scientific research personnel and understand the topology structure and overall function in the field of cooperation network (Liu, 2013).

Community discovery as one of the classic problems of network science, aroused extensive attention of researchers for a long time, and have produced many excellent algorithms, such as KL method based on the separation and algorithm based on modularity, etc. In 2008, Vincent D Blondel puts forward the algorithm of ‘Fast unfolding’ to optimize Modularity value (Meo, Ferrara, Fiumara, Provetti, Flickr & Zhang, et al. 2008), which adopted by Gephi and can be used to detect communities. Gephi was used to generate topological graph of three major communities in virtualization technology research co-author network in China, and there are a total of 182 members in this three communities, accounting for 10.22% of the members on a whole as is shown in figure 5.

![Community 1](image) ![Community 2](image) ![Community 3](image)

**Figure 5 Topological graphs of 3 main communities**

There are 71 nodes and 204 edges in community 1, and the research involved in cloud computing, virtual machine, network virtualization, programmable network. Its members are mainly composed of National University of Defense and Technology (shown in part A),
Chinese Academy of Sciences (shown in part B), Zhejiang University (shown in part C) and Beijing University of Posts and Telecommunications (shown in part D). Wu heng is in the position of cut-vertex, connecting these two subnets of National University of Defense and Technology and Chinese Academy of Sciences. If the node is "lost", community 1 will no longer be connected. Wu heng is a postgraduate of National University of Defense and Technology, who come into the Institute of Computing Technology of the Chinese Academy of Sciences for further study after graduation. Student mobility builds a bridge for communication between the two institutions. Also at the location of cut-vertex is zhang Jianhua and Liu Yunjie. Zhang jianhua connects Zhejiang University to Chinese Academy of Sciences, and Liu Yunjie connects Chinese Academy of Sciences to Beijing University of Posts and Telecommunications. Community 1 consists of the relationship of cooperation between teachers and students, colleagues, researchers in similar area, which has strong cooperative relations of interconnectedness, is conducive to the spread of new knowledge.

Community 2 contains 57 nodes, 136 edges, and studies storage virtualization technology covering cloud computing, such as distributed storage, storage security, disaster recovery, load balancing and optimization algorithm, etc. Members are mainly composed of two subgroups of Huazhong University of Science and Technology (shown in part A) and National University of Defense and Technology (shown in part B). The cooperation between Xiao Nong and Wu Song connected National University of Defense and Technology and Huazhong University of Science and Technology. Jin Hai (the Yangtze river scholar distinguished professor of Huazhong University of Science and Technology, winners for outstanding young scholars), Wu Song (the director of parallel and distributed computing in computer institute, deputy director of service computing technology and system key laboratory of the ministry of education and the cluster and grid computing of key laboratory in Hubei province), Lu Yutong (Tianhe-2 supercomputer system chief designer) is the nodes of larger betweenness centrality, which occupies an important position and has great influence in network, and plays an important role in the expansion of the network in the future.

Community 3 contains 54 nodes, 250 edges, mainly focusing on memory virtualization, network virtualization, virtual network platform and cloud computing, etc. Members include three subgroups of Tsinghua university (shown in part A), The PLA Information Engineering University (shown in part B) and Chinese Academy of Sciences (shown in part C). Li Yongtong belongs to Chinese Academy of Sciences and Tsinghua University, connecting the two subgroups of B and C.

Results, part III: Centrality

"Centrality" index is often used by social network analysis measuring the core role and influence of actors in the network to find out some important or prominent person. In order to evaluate the status and influence of network nodes more scientific and pluralistic, centrality index has been expanded from three indexes of degree centrality (DC), closeness centrality and betweenness centrality (BC) to multiple indexes, such as increasing the indexes of eigenvector centrality (EC), PageRank (PR), AuthorRank (AR), etc.

Based on the idea of "the actor has direct connection with other actors, then the actor is in a central position", degree centrality uses the degree to measure. In scientific cooperation network, higher degree centrality of an author shows that there are more other authors having cooperation with him in publishing articles. However, it does not directly evaluate the importance of the author. As shown in figure 1, the author V3 published an article but degree is three, the author V5 co-authored two articles but degree is 1, so it is difficult to evaluate that the V3 is more important than the V5. Closeness centrality tries to measure the ability that other actors do not control an actor and calculate the distance between a node to another node. In paper cooperation network, if the closeness centrality is relative small, it means that
the author has close cooperative relationship with other authors and the information communication is unobstructed; Otherwise, there is only less cooperation and poor information communication. In a non-connected network, the distance between a node and other nodes may be unpredictable, so the index is only applied to fully connected network. Betweenness centrality measures the control degree of resources for actors. It is generally believed that more times of nodes appear in the geodesic distance of various node pairs, the higher the betweenness centrality, and its control for information exchange and resource flow is more and more important. Betweenness centrality is also applied in the unconnected network, because of many nodes in the network does not serve as a "bridge" role, and thus generates a large number of nodes with zero centrality.

In 1972, Bonacich puts forward that the eigenvector of adjacency matrix maximum eigenvalue can be a good measurement of network centrality degree (Bonacich, 2007). Eigenvector centrality value of a node is determined by the value of adjacent vertices, it not only considers the centrality of adjacent nodes, but also considers the position of the nodes in the network. Eigenvector centrality can distinguish the importance of node in the network of high-degree node and many low-degree nodes or many high-degree nodes and low-degree node, but it can distinguish the difference of nodes clearly in even distributed degree of network. PageRank is Google's core ranking mechanism, if a page has many backward chainings or is linked by the page that has many backward chainings, the page will have a high grade. In virtualization technology co-author network, when the degree of node is higher or the node adjacent to "high quality" node, PR value of the node is also higher. In the undirected co-author network, PageRank calculates the value of nodes by converting undirected edge to directed symmetric edge building binary directed network. However, in the process of converting to binary directed network will lead to the loss of edge weight. Liu X and others propose AuthorRank of using weighted sorting algorithm to solve this problem on the basis of PageRank. The theoretical hypothesis is that the connection weight represents the strength of cooperation between two authors, when edge weight is bigger and the relationship of edges is more important (Liu, 2005). Existing network analysis program is not integrated the sorting algorithm. In this paper, we use Python to realize the algorithm.

From the degree centrality, we can see that Li Yong, Jin Hai and Wu Song are more involved in the publication of papers, namely, the papers that have published having many cooperator. As is known in betweenness centrality, Zhang Jianhua, Wu Heng and Li Yong are in different scientific research group, playing an important role in the network. They are the only bridge connecting those scientific research groups. Although Zhang Jianhua and Wu Heng have little articles, this cannot prevent them from control of the flow of information. If there are no nodes, we will have a sparser network. Li Yong, Zeng Lieguang and Jin Depeng who publish the articles most, have higher eigenvector centrality. It is similar to the ranking results of PR and AR, but is clearly different from the top three centrality ranking result. The top four of AR ranking is Jin Hai, Gou Yudong, Wu Song and Shu jiwu (Tsinghua University) respectively. They have great influence in virtualization technology field, which can cultivate more creative team and lead the development of virtualization technology. In closing we think the ranking results of PR and AR are more accurate combined with the effect of the author. It is shown in table 2.

Table 2 Node centrality of virtualization technology co-author network (Top15)

<table>
<thead>
<tr>
<th>DC</th>
<th>BC</th>
<th>EC</th>
<th>PR</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiYong</td>
<td>ZhangJianHua</td>
<td>LiYong</td>
<td>JinHai</td>
<td>JinHai</td>
</tr>
<tr>
<td>JinHai</td>
<td>WuHeng</td>
<td>ZengLieGuang</td>
<td>GuoYuDong</td>
<td>GuoYuDong</td>
</tr>
<tr>
<td>WuSong</td>
<td>LiYong</td>
<td>JinDePeng</td>
<td>ShuJiWu</td>
<td>WuSong</td>
</tr>
</tbody>
</table>

From the degree centrality, we can see that Li Yong, Jin Hai and Wu Song are more involved in the publication of papers, namely, the papers that have published having many cooperator. As is known in betweenness centrality, Zhang Jianhua, Wu Heng and Li Yong are in different scientific research group, playing an important role in the network. They are the only bridge connecting those scientific research groups. Although Zhang Jianhua and Wu Heng have little articles, this cannot prevent them from control of the flow of information. If there are no nodes, we will have a sparser network. Li Yong, Zeng Lieguang and Jin Depeng who publish the articles most, have higher eigenvector centrality. It is similar to the ranking results of PR and AR, but is clearly different from the top three centrality ranking result. The top four of AR ranking is Jin Hai, Gou Yudong, Wu Song and Shu jiwu (Tsinghua University) respectively. They have great influence in virtualization technology field, which can cultivate more creative team and lead the development of virtualization technology. In closing we think the ranking results of PR and AR are more accurate combined with the effect of the author. It is shown in table 2.

Table 2 Node centrality of virtualization technology co-author network (Top15)

<table>
<thead>
<tr>
<th>DC</th>
<th>BC</th>
<th>EC</th>
<th>PR</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiYong</td>
<td>ZhangJianHua</td>
<td>LiYong</td>
<td>JinHai</td>
<td>JinHai</td>
</tr>
<tr>
<td>JinHai</td>
<td>WuHeng</td>
<td>ZengLieGuang</td>
<td>GuoYuDong</td>
<td>GuoYuDong</td>
</tr>
<tr>
<td>WuSong</td>
<td>LiYong</td>
<td>JinDePeng</td>
<td>ShuJiWu</td>
<td>WuSong</td>
</tr>
</tbody>
</table>
Results, part IV: Small-world effect

Small-world phenomenon is the basic phenomenon of social network, whether from natural system to social system, it can be described by social network model with complex structure and architectural feature of the coexistence of order and randomness (Liu, 2015). It is not easy to establish a mathematical model that can accurately describe the cooperation problem in the society of dynamic change. Watts and Strogatz (1998) proposed the "small world" network model to dynamically reflect the rules and characteristics of network change. Watts points out that the overall changes of network may originated from changes in the local network. For example, in the scientific research cooperation network, deleting a line but adding randomly a line at the same time can significantly change the average path length and clustering coefficient of network.

Small-world network has the feature expanding from near regular networks to completely random networks (Zhichao & Zhu, 2016), whose feature is more in line with the real model of knowledge diffusion. In a scientific cooperation network with a small-world effect, knowledge is easier to spread. The greater access of the individual to gain knowledge diffusion in the same period of evolution, the higher the average level of knowledge in the entire network (Zhichao, et al., 2016). At the same time, it can avoid the possibilities when some researchers have made progress in some scientific research field while other entrants have to start from scratch (Lai, 2012). In addition, there are more potential opportunities for cooperation in small-world network. Researchers get in touch with other researchers and establish the relationship with them, the stronger the small world effect, and the greater the probability of nodes knowing each other (Yang, 2015).

Relative short network diameter and relative large clustering coefficient are two important characteristics of small-world network. Uzzi and Spiro put forward the Q value to evaluate small-world effect (Uzzi, 2005), calculation formula is $Q = \frac{CC_{actual}}{CC_{random}} \frac{PL_{actual}}{PL_{random}}$, Where $CC_{actual}$ is the clustering coefficient for observation network and random network, $APL_{actual}$ is average path length for observation network, and $PL_{random}$ is average path length of the random network. If Q is much higher than 1, it indicates that observation network has the features of small world. Humphries (2008) and Liu (2015) have made similar conclusions through research.

According to calculation formula of Uzzi, this paper used IGraph package of R platform. We carried out 1000 simulations that take 1780 nodes, 4 adjacent point number (average degree is...
3.934), 2 probability of rewiring as parameter. Then, we calculate the average clustering coefficient and average path length distribution of random network through the results of simulation and use the Q value to evaluate the small-world effect of virtualization technology research co-author network. As is shown in figure 6 and figure 7, we know that the scope of Q value is [5.362, 6.435], far higher than 1. Thus, we can conclude that virtualization technology research co-author network of china has small-world effect, and the small-world effect is more prominent.

```r
ntrials=1000
nv=1780
ad=4
cl.rg =numeric(ntrials)
apl.rg=numeric(ntrials)
for(i in (1:ntrials)){
g.rg=watts.strogatz.game(1,nv,ad,0.2)
cl.rg[i]=transitivity(g.rg,type = "average")
apl.rg[i]=average.path.length(g.rg)
}
# Calculate the range of Q values
cl.actual=0.983
apl.actual=3.934
Q =( cl.actual /cl.rg)/( apl.actual /apl.rg)
> summary(Q)
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     
5.362  5.854  5.949  5.955  6.055  6.435
```

Figure 5 Small-world effect simulation results of virtualization technology research co-author network

Figure 6 Scatter diagram of virtualization technology research co-author network Q value distribution

Results, part V: scale-free phenomenon analysis

Social network often has a small number of vertices that have a very higher degree, the rest are vertices of low degree, and this phenomenon is commonly known as the "Matthew effect". Scientific research cooperation in agency team and teacher-student team is a common way of scientific research cooperation. In addition, general researchers tend to establish scientific research relationship with the researchers who have strong scientific research ability. In scientific research cooperation network, famous scholars, team leader or postgraduate tutor usually have a high degree and other person have a low degree. The network where its degree has no obvious characteristic length is called the scale-free network. The whole scale-free networks connectivity is dominated by a small number of core nodes that have heavily connected, once it is attacked and then collapsed will lead to paralysis of the entire network.
However, there is a little influence on the stability of the network when it randomly attacks other nodes in the network, so "robustness and vulnerability" is one of the main characteristics of scale-free networks. Whether virtualization technology research co-author network has scale-free properties worth researching and analysis.

Tail on the right side of scale-free network degree distribution follows power law: if a degree distribution in the network can be expressed with power law distribution function $p(k) = Ck^{-\alpha}$ ($C$ is the normalization constant, $\alpha$ is a power value, and $\alpha \in (2 \sim 3)$), then it can be considered that the network has scale-free properties (Nooy, et al. 2012).

Scholars commonly use the way of goodness of fit test to detect whether the data distribution has certain properties. Also, the way of goodness of fit test were used to study the degree distribution of virtualization technology research co-author network whether has power-law distribution property or other properties. There are many types of statistic of goodness-of-fit; KS statistic is relatively commonly used. Under the hypothesis of data is the result of the power-law distribution; Clauset recommended that adopting the bootstrap statistical method by sampling repeatedly to carry on the goodness of fit test (Clauset, Shalizi & Newman, 2009). The method by calculating artificial data $K_5^*$ statistic is greater than the KS statistic $p$ ratio of the distance between the cumulative distribution of the observed data and the fitting power law distribution model (Yang, 2014). If the $p$ value is relative high, the assumption that the distribution of data is the result of the power law is compatible, but it cannot be ruled out whether observation data remains subject to Poisson Distribution, log-normal distributed, exponential distribution and other distribution; When it needs to rule out other distribution possibilities, we can use the same method again. If the $p$ value is very small, the power-law distribution cannot reasonably fit the observation data, choosing another distribution may be more appropriate. Clauset thinks that it is appropriate to choose 0.1 as threshold value of $p$, namely, $p > 0.1$, which the hypothesis can be considered as valid and the converse is not true. In the case of observation data collection subject to multiple distributions, likelihood ratio test is also commonly used in test. It chooses a more representative observation data model by comparing two distributions of likelihood function or log-likelihood ratios. Vuong improved the calculation method of the likelihood ratio test and gave the $p$ value that can judge a model whether good or not (Vuong, 1989). Generally, it is more appropriate for $p$ to choose 0.1 as the value.

PoweRlaw package of R contains a number of functions and methods, which can fit the sample data whether is derived from the distribution of power law, lognormal, Poisson, exponential and have a test, as well as compare model and give the $p$ value that can evaluate whether the model is good or not (Gillespie, 2014). This article uses the poweRlaw package to research the degree distribution of virtualization technology co-author network. Figure 8 and figure 9 assumed that virtualization technology research co-author network degree distribution are power-law distribution, lognormal distribution, Poisson distribution, exponential distribution and have the process and result of goodness-of-fit test calculation; Test of $p$ values were 0.004, 0.473, 0.0005, 0.667. Therefore, there is little possibility that the network degree distribution follows power-law distribution and Poisson distribution, and lognormal distribution or exponential distribution fit well. Which distribution are more suitable needs to be determined by the likelihood ratio test. Figure 10 shows the calculating process and results of the likelihood ratio test. Through the results of single tail test, the lognormal curve can fit the distribution of degree of virtualization technology research co-author network.

```r
# Read Data File
degree=read.table(file.choose())
degree=as.integer(degree$V1)
# Load Package
library("poweRlaw")
# Create power-law distribution object, using 2000 bootstrap simulation fitting
```
virtual_pl <- displ$new(degree)
virtual_pl$setXmin(estimate_xmin(virtual_pl))
bs_virtual_pl = bootstrap_p(virtual_pl, no_of_sims=2000, threads=10)
# Create log-normal distribution object, using 2000 bootstrap simulation fitting
virtual_ln <- dislnorm$new(degree)
virtual_ln$setXmin(estimate_xmin(virtual_ln))
bs_virtual_ln = bootstrap_p(virtual_ln, no_of_sims=2000, threads=10)
# Create Poisson distribution object, using 2000 bootstrap simulation fitting
virtual_po <- dispois$new(degree)
virtual_po$setXmin(estimate_xmin(virtual_po))
bs_virtual_po = bootstrap_p(virtual_po, no_of_sims=2000, threads=10)
# Create exponential distribution object, using 2000 bootstrap simulation fitting
virtual_exp <- disexp$new(degree)
virtual_exp$setXmin(estimate_xmin(virtual_exp))
bs_virtual_exp = bootstrap_p(virtual_exp, no_of_sims=2000, threads=10)

Figure 7 virtualization technology research co-author network degree distribution fitting process

Figure 8 cumulative distribution of virtualization technology research co-author network degree. (the red line is the power law fitting, the blue line is lognormal fitting, the green line is Poisson fitting, the black line is exponential fitting)

Figure 9 process and results of likelihood ratio test

Discussion and Conclusions
By analyzing the virtualization technology research co-author network in China, we found some valuable phenomenon, characteristics and law in the field of the technology research.

(1) The scale of virtualization technology research co-author network in China is small, with sparse overall connection, many components and high modularity, but small in larger community. There is more small research groups in virtualization technology research area and is closely linked in internal cooperation of the groups, but very loose relationship between groups, reflecting the diversity of virtualization technology research.

(2) Three members of co-author community in the fields of virtualization technology research in China are mainly from Chinese Academy of Sciences, National University of Defense and Technology, Huazhong University of Science and Technology, Tsinghua University and other institutions. In the information communication network of scientific research cooperation,
close cooperation between groups is the basis of an academic group closely focusing on a particular direction for centralized breakthrough. Usually, small groups are based on the mutual interest and complementary field of peer experts. These small groups are the bridge and the link of knowledge activities in different disciplines, regions and organizations, which plays a very important role in promoting knowledge diffusion, knowledge transfer and knowledge spillover. If we give full attention to these small groups in the scientific research management activities, it can make scientific research management more scientific, more targeted, and management effect will be more prominent (Yang, 2015).

(3) Scientific research cooperation is usually divided into five types: employers and employees cooperation, cooperation between teachers and students, cooperation between researchers in similar areas, cooperation between researchers in different areas and the cooperation of different gender (Xie, 2006). From the community analysis of virtualization technology research co-author network in China, you can see that the cooperation between teachers and students, cooperation between the researchers in similar areas and the cooperation between the researchers in different areas are the main types of cooperation.

(4) In China’s virtualization technology research co-author network, there are two kinds of important members: one is nodes of higher AR value, such as Jin Hai, Guo Yudong, Wu Song and Shu Jiwu, etc. These scholars with high influence are usually the leader in various research institutions or research field; another one is a bridge among the groups, such as Zhang Jianhua, Wu Heng, Li Yong, Wu Yuanquan, etc. They are the main channels to obtain external knowledge and resources in the same group, also the key channel of knowledge diffusion and transfer in different groups. In terms of scientific research management, we should pay attention to these academic leaders and scholars in the "cut-point" position.

(5) Virtualization technology research co-author network in China has stronger small-world effect, the researchers can efficiently communicate in cooperation with related researchers and can easily access to the new knowledge with other research groups. Knowledge is easier to spread in the network; average knowledge level of the entire network is high. At the same time, virtualization technology research area in our country contains more cooperative opportunities, As time goes by, originally loose network may be more coherent.

(6) Different from scale-free properties of co-author network that studied by the scholars of Huang Kaimu (2015), Zhang (2012) and Meng (2009), the paper found that virtualization technology research co-author network degree distribution is unable to fit power law curve but fit well with lognormal curve through simulation. Is the cause of the data itself or caused by the characteristics of the research area, it remains to be studied.

References


A Weighted Method for Citation Network Community Detection

Yunwei Chen*, Xue Xiao, Yong Deng, Zhiqiang Zhang

* Corresponding Author: chenyw@clas.ac.cn
Scientometrics & Evaluation Research Center (SERC), Chengdu Library and Information Center of Chinese Academy of Sciences, Chengdu, 610041 (China)

Abstract
There are many reasons for studying the community structure in citation network, whose underlying community structure may help us understand both the obvious and the more subtle interrelations between subfields, which are so important for understand the research structure and paradigm structure of a subject. However, most methods used for citation network community detection are based on topological structure or content separately. Usually such link-based methods ignore the content of citations, and the content-based methods lack the consideration of citation networks’ topological structure, which might be crucial for accurate and meaningful community detection. In this paper, we develop a weighted method to identify the community structure of citation network. The new weighted model incorporates both the topological links and content information in citation networks to find communities. A text similarity method is used for weighting the links in citation networks and then constructs a weighted citation network to carry out community detection analysis. Two case studies had been carried out in two fields of Scientometrics and Evolving Citation Networks. Results revealed that using weighted citation edges with VSM-based nodes similarity to detect communities can improve the modularity Q value, which might mean a better assignment result. Four nodes’ movement with the direction of better appropriate of community assignation in the 188 node citation network of the field Scientometrics had also indicates the advantage of our weighted method.

Conference Topic
Social network analysis.

Introduction
In this work, we develop a weighted method to identify the community structure of citation network. There are many a kind of reasons for studying the community structure in citation network, whose underlying community structure may help us understand both the obvious and the more subtle interrelations between subfields, as well as the growth and the ebb of subfields (Chen & Redner, 2010).

Communities are the clusters of closely connected nodes within a network, which is a property found in many networks, such as biological networks, the World Wide Web, social networks, collaboration works, citation networks, et al (Girvan & Newman, 2002). To identify communities within networks, a variety of methods have been developed. A recent new and powerful method to detect communities in complex networks is the study of Girvan and Newman (2002), who boomed the community research for their new algorithm using information about edge betweenness (Newman, 2001) to detect community peripheries, we call GN algorithm. However, the GN algorithm and some other algorithms are computationally demanding, which limits their application only to small networks. Therefore, Newman (2004a) proposed a new much faster algorithm, typically thousands of times faster than previous algorithms, we call FN algorithm. FN operates on different principles to GN but gives qualitatively similar results. FN algorithm is based on the idea of modularity, with the optimization of Q values over all possible divisions to find the best communities. Modularity measure Q is calculated to know when the communities found by the algorithm are good ones (Newman & Girvan, 2004). It has become a popular approach of optimizing a modularity function to community detection (Waltman & Van Eck, 2013). Compared to the earlier community detection methods, Q requires no extra knowledge beyond the network structure itself and can be applied to any type of network (Chen & Redner, 2010).
Nevertheless, the above algorithms proposed in the literature are suitable only for small and medium-sized networks. Therefore, some optimization algorithms were introduced to produce high-quality results even for very large networks, such as citation networks with tens of millions of nodes and hundreds of millions of edges. The first existing algorithms for large-scale modularity optimization is proposed by Blondel et al. (2008), often referred to as the Louvain algorithm. The other one is Louvain algorithm with multilevel refinement proposed by Rotta and Noack (2011), which is an extension of Louvain algorithm. Built on the ideas from these two existing algorithms, Waltman & Van Eck (2013) proposed a smart local moving (SLM) algorithm. All the three algorithms rely heavily on a well-known local moving heuristic (LMH), the idea of which is to repeatedly move individual nodes from one community to another in such a way that each node movement results in a modularity increase. We do not want to discuss the details of these algorithms for the purpose of this paper is not to reconstruct or modify them, but to make more meaningful citation networks and provide for the algorithms to detect communities. The theories and steps of the algorithm could be found in these papers: Girvan & Newman (2002), Newman (2004a), Newman & Girvan (2004), Waltman & Van Eck (2013).

Benefitted from the progress of community detection algorithms recent years especially as discussed above, lots of works focusing on the community structure of many kinds of networks had been carried out, which includes citation networks.

In the next section, we outline the principle of modularity measure \( Q \), the Louvain algorithm that we use to analyse community structure within citation network, and the related works about citation networks community detection.

Based on the below related works analysis, we found that although many methods had been used to disclose the community structure of citation networks, there were still very few works focused the difference of different citation links. Therefore, this study advanced the existing works by constructing a weighted citation network to carry out community detection analysis. To our knowledge, none of the existing studies has done such work as ever. The significance of this work is that using weighted citation edges with VSM-based nodes similarity to detect communities with an improvement value of modularity \( Q \), which might mean a better assignment result. Our weighted method breaks the barriers of traditional citation network analysis that regarding all edges as the same importance. We find that which avoids some of the shortcomings of the traditional techniques. It would be very valuable for many kinds of scientists, policy makers and stakeholders to deal with it.

Related works

In this section we review some representative existing methods for detecting communities and make a brief introduction to their advantages and disadvantages.

Modularity measure \( Q \)

Modularity \( Q \) is the fraction of edges that fall within communities, minus the expected value of the same quantity if edges fall at random without regard for the community structure. If a particular division gives no more within-community edges than would be expected by random chance we will get \( Q = 0 \). Values other than 0 indicate deviations from randomness, and in practice values greater than about 0.3 appear to indicate significant community structure (Newman, 2004a).

Newman & Girvan (2004) defined \( Q \) as follows,

\[
Q = \sum_i (f_{ii} - a_i^2)
\]

where \( f \) is a \( k \times k \) symmetric matrix whose element \( f_{ij} \) is the fraction of all edges in the network that link vertices in community \( i \) to vertices in community \( j \). So they further define the row (or
column) sums $a_i = \sum_j f_{ij}$, which represent the fraction of edges that connect to vertices in community $i$. In a network in which edges fall between vertices without regard for the communities they belong to, we would have $f_{ij} = a_ia_j$.

Newman (2004b) defined $Q$ in weighted networks (edges have weights, for example, based on the number of co-authored papers between two authors) as:

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{K_iK_j}{2m}] \delta(c_i, c_j)$$

where $c_i$ denotes the community to which node $i$ has been assigned, $A_{ij}$ represents whether there is an edge between nodes $i$ and $j$ ($A_{ij} = 1$) or not ($A_{ij} = 0$) in un-weighted networks, or the weight of the edge between node $i$ and $j$ in weighted networks. $K_i$ is the degree of node $I$ in un-weighted networks or the sum of the weights of the edges linked to node $i$, the $\delta(c_i, c_j)$ indicates whether nodes $i$ and $j$ belong to the same community. It equals 1 if $c_i = c_j$ and 0 otherwise. And

$$m = \frac{1}{2} \sum_{ij} A_{ij}$$

is the total number of edges in the network.

There is also a third formula of $Q$ can be written as:

$$Q = \sum_{i=1}^{n} \left[ \frac{l_i}{L} - \frac{d_i}{2L} \right]^2$$

Here $n$ is the number of modules, $L$ is the number of links in the network, $l_v$ is the number of links between nodes in module $v$, and $d_v$ is the sum of the degrees of the nodes in module $v$.

**Louvain algorithm**

Louvain algorithm is proposed by Blondel et al. (2008), which finds high modularity partitions of large networks in short time and that unfolds a complete hierarchical community structure for the network, thereby giving access to different resolutions of community detection. This algorithm is based on the idea of the local moving heuristic is to repeatedly move individual nodes from one community to another in such a way that each node movement results in a modularity increase. It is divided in two phases that are repeated iteratively. Assume a network with $n$ nodes. First, we begin with each one node is a community. The algorithm then uses the local moving heuristic to obtain an improved community structure. Hence, in this initial partition individual nodes are moved from one community to another until no further increase in modularity can be achieved. At this point, a reduced network is constructed. Second, we build a new network whose nodes are the communities found during the first phase. The weights of the links between the new nodes are given by the sum of the weight of the links between nodes in the corresponding two communities. Once this second phase is completed, it is then possible to reapply the first phase of the algorithm to the resulting weighted network and to iterate until there are no more changes and a maximum of modularity is attained (Blondel et al., 2008).

**Communities in citation networks**

Citation network is originated from the reference relationships. There are many researches show that there is obvious trend of concentration and discrete in citation network, and publications in the same field seems like to cite semblable publications. Community in the citation network represent related papers on a single or some closely relevant topics. Community detection method is used for finding related papers in a given field. We could find the communities in citation networks and help researchers or policy makers to watch the whole pattern of a field at a glance. Subelj et al. (2016) presented a systematic comparison of the performance of a large number of citation networks clustering methods and found that
there is a trade-off between different properties that may be considered desirable for a good clustering of publications. At present, the methods used for citation network community detection could be divided into two classes. One is based on topological structure and another is based on content, such as text similarity.

Topological structure based methods usually use the community detection algorithms in the field of complex network to carry out empirical study in citation networks. For example, Wanjantuk et al. (2004) used the random walk graph clustering algorithm basing only on the link structure of a graph that can efficiently identify highly topically related communities to identify the communities within citation graph. Kajikawa et al. (2008) tracked the emerging research domains in energy research by using citation network analysis. They used FN algorithm to divide the largest connected component into some communities. Chen & Redner (2010) investigated the community structure of physics subfields in the citation network of all Physical Review publications between 1893 and August 2007. They focused only on well-cited publications that received more than 100 citations, and applied modularity maximization to uncover major communities that correspond to clearly identifiable subfields of physics. Ren et al. (2012) proposed a new clustering model for the high clustering in citation networks and got higher clustering coefficient and the size distribution of co-citation clusters in real networks. Wu et al. (2015) developed a local community detecting algorithm to find corresponding research group. Kusumastuti et al. (2016) identified the clusters and sub-clusters of citation networks using the CitNetExplorer software to visualize timeline-based citation patterns.

Content based methods usually use traditional clustering algorithms to classify citation networks when the virtual links basing similarity are used to replace the actual links. For instance, Aljaber et al. (2010) presented a new approach for clustering scientific documents basing on the utilization of citation contexts. Chen et al. (2013) proposed a community discovery algorithm of citation semantic link network to find the semantic community in citation semantic link network. Liu et al. (2015) proposed a citation similarity based community detection method by transforming citation network to paper similarity network to detect the communities in citation networks. Comparing the above two kinds of methods we could find that topological structure based methods usually lack the consideration of content-based text similarity, while the latter methods have also ignored the networks’ topological structure. Nevertheless, both the above two kinds of methods could not fulfil the need of citation network community analysis when used separately. Therefore, there are also some studies had paid attention to this question. For example, Cohn & Hofmann (2001) described a joint probabilistic model for modelling the contents and inter-connectivity of document collections such as sets of research paper archives. Erosheva et al. (2004) explored an internal soft-classification structure of articles based only on semantic decompositions of abstracts and bibliographies and compared it with the formal discipline classifications. Pathak et al. (2008) presented a Bayesian generative model for community extraction which incorporates both the link and content information in networks and its empirical study had been carried out on an email network. Although these papers did not analyse the citation network’s community, they indeed provided a powerful idea for community detection in citation networks. Indeed, most such studies incorporating both the topological link and content information had been studied in traditional complex networks, such as.

Therefore, this study advanced the existing works by incorporating both the topological links and content information in citation networks to find communities. A text similarity method is used for weighting the links in citation networks and then constructs a weighted citation network to carry out community detection analysis.
Data and methods

Data

This paper prepared two datasets to carry out case study. One is all types of 4417 publications published in the journal of *Scientometrics* in 1978–2016 (publication year) were downloaded and chosen as base data of case study construct citation network. Data was acquired from the Web of Science in March 13 2016. The citation network consists of 1483 publications published in *Scientometrics* between 1978 through January 2016 with at least one citation (the nodes) to the other articles among the 4417 publications. At the same time, publications have no any field of title, abstract and keywords had been removed from the network. Therefore, the 1483 nodes make up a network with 28 weakly connected components through 3072 arcs. Among which the largest connected component consists of 1420 nodes, which had been used to find communities according to different weighted methods or unweighted method.

The other is the 2046 publications (article, letter, proceedings paper and review) of Evolving Citation Networks published before 2014 in WOS. Search terms to retrieval papers are: TS=((citation* and (evolv* OR evolut* OR "knowledge flow" OR path OR trajector* OR backbone)) NOT "citation classic"). The citation network consists of 340 publications with at least one citation (the nodes) to the other articles and also has all fields of title, abstract and keywords among the 2046 publications. Therefore, the 340 nodes make up a network with 32 weakly connected components through 531 arcs. Among which the largest connected component consists of 246 nodes, which had been used to find communities according to different weighted methods or unweighted method.

Although both the citation networks have more than one connected component, most nodes locate in the largest connected component. In this paper, only the largest connected components are included. On the other hand, the directed arcs in citation networks are treated as undirected edges used for community detection in this paper. Science of Science (Sci²) Tool (http://cns.iu.indiana.edu) and Gephi were used to citation network construction and community detection.

Weighted method- VSM-based nodes similarity calculation

The Vector space model (VSM) is a mathematical model first presented by Salton, Wong and Yang (1975) and generally defined by Salton (1989) as an algebraic model for representing text documents as vectors of identifiers such as index terms. It is now popular used in information filtering, information retrieval, indexing and relevancy rankings.

In VSM documents are represented as mutually independent ordered term vectors, \((T_1, T_2, T_3, \ldots, T_n)\). Each term \(T_i\) has its own weight \(W_i\) according to the contribution to the documents. Thus document \(d_j\) is represented as vector \(d_j = (W_{1,j}, W_{2,j}, W_{3,j}, \ldots, W_{n,j})\). Each dimension corresponds to a separate term. The value of a term \(i\) in the vector is non-zero if it occurs in the document. Then we use the TF-IDF weighting method to computing the values of \(W_i\). However, the TF-IDF algorithm does not consider the position of a term in a document. There are some studies show that the importance of a term is different according to the position in the document, such as title, keywords and abstract. Therefore, three groups of position weight \(\gamma\) are added to \(W_i\) based on the following formula.

<table>
<thead>
<tr>
<th>(\gamma)</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>2.5</td>
<td>1.5</td>
<td>When term (i) lies in keywords</td>
</tr>
<tr>
<td>1.0</td>
<td>1.5</td>
<td>2.5</td>
<td>When term (i) lies in title</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>When term (i) lies in abstract</td>
</tr>
</tbody>
</table>

Thus, the weight of term \(i\) is calculated as follows.

\[ W_i = \text{TF}(i) \times \text{IDF}(i) \times \gamma(i) \]
In the case of a term appears in more than one position in a document, the highest value of $\gamma$ is accepted.

Finally, we use cosine similarity to calculate the similarity between a pair of vectors. It has to be noted that although citation links have directions, the original citation networks that we use are regarded as un-weighted and undirected and do not have loops. In our citation networks the cosine similarity between two linked documents is regarded as the weight of the link.

**Results and Analysis**

We have used both the two largest connected components of the two datasets to perform basic tests of the significance of the communities that are discussed in this work based on the three group of $\gamma$. First, we performed the same community detection analysis described above for the two set of the two largest connected components without weights. Second, VSM based weighted method had been accepted to reconstruct citation network at the three ways of $\gamma$. Results of community detection in the two datasets are listed in table 1.

<table>
<thead>
<tr>
<th></th>
<th># communities</th>
<th>modularity Q</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scientometrics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(largest connected component consists of 1420 nodes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unweighted</td>
<td>20</td>
<td>0.672</td>
</tr>
<tr>
<td>weighted group 1</td>
<td>26</td>
<td>0.764</td>
</tr>
<tr>
<td>weighted group 2</td>
<td>30</td>
<td>0.780</td>
</tr>
<tr>
<td>weighted group 3</td>
<td>28</td>
<td>0.769</td>
</tr>
<tr>
<td><strong>Evolving Citation Networks</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(largest connected component consists of 340 nodes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unweighted</td>
<td>13</td>
<td>0.722</td>
</tr>
<tr>
<td>weighted group 1</td>
<td>14</td>
<td>0.743</td>
</tr>
<tr>
<td>weighted group 2</td>
<td>14</td>
<td>0.730</td>
</tr>
<tr>
<td>weighted group 3</td>
<td>12</td>
<td>0.739</td>
</tr>
</tbody>
</table>

**Notes:** Calculated by Gelphi, resolution (Lambiotte et al., 2009): 1.0, randomize: on. Communities detection method is based on Blondel, et al.(2008).

Table 1 shows the increasing values of modularity Q through weighted method for community detection in both case of Scientometrics and Evolving Citation Networks no matter which $\gamma$ was used. However, the result could not show which group of $\gamma$ is a better choice to improve the performance of community detection. For example, $\gamma$ group 2 seems to be a better choice for the case of Scientometrics but it looks like the worse for the case of Evolving Citation Networks, for which $\gamma$ group 1 might be better. The reason might be related to the particular feature of a field or the size of a network.

Thus what we can conclude is that our weighted method truly improved the ability to find communities in citation networks when evaluated by the values of modularity Q. In order to find what this difference is, we decided to investigate further the difference between weighted and unweighted methods. Weighted method with $\gamma$ group 2 was used to be carry out comparison analysis to unweighted network. Meanwhile, in order to keep the scope manageable to get a more careful look at the nodes moving between weighted and unweighted methods, we restrict ourselves to well-cited 188 publications by k-core >4 in the 1420 nodes of the case of Scientometrics, defined as those with higher degrees, to be shown in visualization maps. Then we got five communities. We present the topics of the five communities by co-keywords in Table 2, which are patent analysis, collaboration, research evaluation, university evaluation and science mapping.
Table 2. Topics of communities in Scientometrics dataset

<table>
<thead>
<tr>
<th>Communities</th>
<th>Topics</th>
<th>keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Patent analysis</td>
<td>Patent citation, Patent mining</td>
</tr>
<tr>
<td>2</td>
<td>Collaboration</td>
<td>Scientific collaboration, Collaboration network</td>
</tr>
<tr>
<td>3</td>
<td>Research evaluation</td>
<td>Research evaluation, h-index</td>
</tr>
<tr>
<td>4</td>
<td>University evaluation</td>
<td>Arwu, Ranking of universities</td>
</tr>
<tr>
<td>5</td>
<td>Science mapping</td>
<td>Mapping of science</td>
</tr>
</tbody>
</table>

Fig.1 illustrates the communities of the 188 publications in both unweighted and weighted networks, which shows that four nodes move between the four networks. They are n46, n1116, n686 and n1386. Their titles are listed in table 3. In order to judge whether the move of the four nodes from unweighted network to weighted network is positive or not, we read the full texts of the four publications carefully and summarized the themes comparing to the topics of the five communities.

Table 3. Four nodes distributed in different communities in Scientometrics dataset

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Title</th>
<th>Communities</th>
</tr>
</thead>
<tbody>
<tr>
<td>n46</td>
<td>Knowledge network centrality, formal rank and research performance: evidence for curvilinear and interaction effects</td>
<td>3</td>
</tr>
<tr>
<td>n686</td>
<td>Bibliometrics evaluation of research performance in pharmacology/pharmacy: China relative to ten representative countries</td>
<td>5</td>
</tr>
<tr>
<td>n1386</td>
<td>The 100 most prolific economists using the p-index</td>
<td>5</td>
</tr>
</tbody>
</table>

Figure 1. Comparison of communities in Scientometrics
Publication n46 (Badar, et al., 2015) explored the curvilinear association of co-authorship network centrality, degree, closeness and betweenness and the research performance, which considered formal rank of the authors as a moderator between network centrality and research performance. Its theme relates to both the topics of community 2 (collaboration) and community 3 (research evaluation). But by reading the paper we found that the central theme of n46 is scientific collaboration network analysis, and the usage for research evaluation is just an extended function. Therefore, assigning n46 to community 2 might be a better choice.

Node 1386 (Prathap, 2010) developed a new indicator called performance index (p-index), which was used to rank 100 most prolific economists. Node 686 (Ding, et al., 2013) evaluated the productivity of China in the field of pharmacology/pharmacy during 2001-2010 comparing to ten representative countries. Both of these two papers had made some mapping analysis but the key purposes were to carry out evaluation analysis. Therefore, these two publications were assigned preferably to community 3 by our weighted method.

Node 1116 (Zhao & Zhang, 2011) used the methods of co-word analysis, social network analysis and mapping knowledge domains as theory basis to construct the co-word network of the field of digital library research. It presented the study status quo and the research paradigm structure of digital library in China. Although it had also made some comparative analysis to that of international digital libraries research, the main contribution is illustration by mappings but not evaluation. Therefore, on the contrary to n1386 and n 686, n1116 was moved to community 5 from community 3 by our weighted classification method.

The above analysis indicates that four nodes all move with the direction of better appropriate of community assignation through our weighted method. In fact, the four nodes are only key nodes with highly degrees among 188 nodes. There are much many nodes with lower degrees surrounded with the four moved nodes will also move with them to new communities in process of our new weighted method. In other words, our new weighted method will rebuild the research structure and paradigm structure of a subject, such as Scientometrics in this case. It might contribute to the better insight into the development of the field of Scientometrics.

Conclusion

Our study provides initial insights on the weighted method for community detection in citation network. An interesting advantage of using weighted community detection method is the fact that breaking the barriers of traditional citation network analysis that regarding all edges as the same importance. We find that using weighted citation edges with VSM-based nodes similarity to detect communities can improve the modularity Q value, which might mean a better assignment result. Four nodes’ movement with the direction of better appropriate of community assignation in the 188 node citation network of the field Scientometrics had also indicates the advantage of our weighted method.

However, by the result of this paper, we are not sure which part of a paper contributes more to the theme similarity between papers. Thus, an important question for further research is whether there is a common weights distribution feature among different parts of a paper or not, for example title, keywords, abstract and even full text.

Another problem with the citation network is that citation is strictly ordering in time and there is a problem of publication delay about months to years from a manuscript’s submission to publication or online when other authors can read it then cite it. As a result, a lot of highly relevant papers did not cite each other. That is to say, some correlative papers have no links in citation networks and suppressed the function of community detection in citation network to
depict the research structure and paradigm structure of a subject. Under this situation such kinds of highly relevant papers might be co-cited by later publications or have relation of bibliographic coupling by citing at least one common earlier publication. Thus, we suggest such publication pairs should be constructed as virtual edges and also be added selectively to our citation network. So we are facing quite a disturbing puzzle: how many or what kinds of such pairs need to be added to our network? We will take this question as one of our future research works.

We also hope that future research will further examine this question in line with deeper investigation into much more factors that could affect citation behaviour, such as co-author, co-institution, even the bias on language, institutions or countries.

Beyond all the problems identified above, we should also pay close attention to the dynamic evolving patterns of citation network, particularly on the transition of communities including merging, dividing, expanding, et al. For instance, Jung et al. (2014) proposed methods to analyse how communities change over time in the citation network graph without additional external information and based on node and link prediction and community detection.

Acknowledgements
This work is funded by the Documentation and Information Special Project of Chinese Academy of Sciences (2016). This work is supported in part by the West Light Foundation of the Chinese Academy of Sciences, China under grant no. [2013]165(3-6).

References
Liu, T.P. & Li, K. (2015). A Citation Similarity Based Community Detection Method in Citation Networks. 2015 Ieee Advanced Information Technology, Electronic and Automation Control Conference (Iaeac), 146-149.


Research on Domain Knowledge Network Based on Bibliometrics

Mingkun Wei¹* Rongying Zhao² Danyang Li³

¹ weimingkun24@163.com
² zhaorongying@126.com
³ whusimldy@163.com

Research Center for Chinese Science Evaluation, Wuhan University, Wuhan, 430072, China
School of Information Management, Wuhan University, Wuhan, 430072, China

Abstract

Knowledge network is widely applied in such fields as management science, library science, information science and pedagogy to determine the cutting-edge and popular aspects in relevant fields. By drawing the knowledge network, the theoretical system of the subject or relevant fields can be constructed and the coordination relationship between the author or organization and the nation can be analyzed. This paper proposes realizing process of the domain knowledge network based on Bibliometrics after defining knowledge network and the basic concepts. The domain knowledge network model is constructed from time-dimension, network type and research field. This research is helpful to construct and analyze domain knowledge network from different levels of knowledge unit and knowledge relation.

Conference topic

Knowledge discovery and data mining

Introduction

Problems and methods are the basic elements of any academic research. The appropriate research method is the key to solve the problem. We found that there were many disciplines and academic fields to identify the research frontier and hotspot by drawing knowledge network in recent years. Studies on the knowledge network put emphasis on the following points: ① Study on the knowledge network structure, behaviors and complexity from the perspectives of complex network, collaboration network and social network. ② Study on the knowledge network composition and activities from the perspective of knowledge management. ③ Explanation and application of the knowledge network theories in the cluster from the perspective of enterprise cluster. ④ Study on the impact of the knowledge network on the improvement of innovation capability and performance in combination with technical innovation.⑤
Study on the information service applications of the knowledge network in the Internet and digital library fields from the perspective of information technology application. It can help the researchers to construct the theoretical system of subject, to measure the cooperative relationship between authors or organizations and states, to explore issues of disciplinary paradigm. Knowledge networks have been widely used in different disciplines, such as management, library and information science, education, etc. It has become the most widely used research method to research the related issues by drawing the knowledge network in the field of Bibliometric, literature analysis, literature review, etc. in addition to frequency statistical methods. There are many scholars cannot make good use of it, though more and more scholars are beginning to use the knowledge network. Therefore, it is necessary to sort out and summarize for the method of knowledge network from the height of methodology. The paper uses CiteSpace to analyze the keywords, references, and category for visualizing domain knowledge network. CiteSpace is a Java-based application that research out by Chen Chaomei (Chen, 2006), and then the paper studies the related concepts, theoretical foundation, drawing tools, network measure, etc. It can help scholars to understand accurately.

The concept of knowledge network was firstly proposed by the Swedish industry but has not been clearly defined yet. Here are some of the most representative ideas on the concept of the knowledge network: It is an aggregate that gathers academic experts, information, and knowledge for the purpose of analysis on a special issue; it is a dynamic framework that consists of the behavioral agent, relationship, and resource system features for value creation through knowledge creation and transfer; and it is a network formed by connection among bodies of knowledge, such as people and enterprises. Recent year’s knowledge network had applied in social analysis and natural sciences (Sonnenwald and Wildemuth et al., 2001; Newman and Barabasi et al., 2006; Hall, 2010). The concept of knowledge network is divided into following level, which includes engineering electrical electronic, computer science and artificial intelligence, management, information science and library science, pedagogy, etc. In the field of management, it mainly refers to the knowledge creation, transfer, use and dissemination of the internal and external enterprises. The knowledge network can be divided into individual network, group network, organization network, enterprise alliance network, business association network, etc. from the perspective of management; the knowledge network is generally refers to the semantic network, concept network, neural network, etc. from the perspective of computer science and artificial intelligence; knowledge network is generally a kind of learning method or tool from the perspective of pedagogy; it mainly refers to the knowledge node and its structure, which is used to reflect the relationship among knowledge organization, storage, retrieval, transfer and utilization in scientific research activities from the perspective of information science and library science. The paper mainly studies the knowledge network in the field of library and information science, so the meaning of the knowledge network that appears in the following paper is belonging to the field of Bibliometric.

The knowledge network coupled with Bibliometric is often used to study collaboration networks. On the knowledge network, vertex generally indicates the storage unit for a knowledge element according to certain rules and principles (Hai, 2006). It can be a book, essay, patent, information fragment or word according to the
size. An edge is the relationship between knowledge elements, represented by the citation relationship on the citation network or word network and the co-occurrence relationship on the co-occurrence network (such as the co-word, co-authoring, and cooperation networks. Therefore, the knowledge network method based on Bibliometric can be defined as ‘it explores the research front and hotspot of subject or related disciplines by drawing different kinds of knowledge network, and it can also construct the theoretical system of discipline or analyze the collaboration between the author, organization, state, etc. which is the research method to explore the discipline paradigm.’ In a variety of specific research literature, knowledge network method is commonly used in visual map, scientific knowledge map, co-word analysis, co-occurrence analysis, etc. But it has special meaning to knowledge network method based on Bibliometric in this paper, which should have three conditions, the data is from the document database, showing by the visual map, the knowledge network is composed by nodes and tie.

Method

1). Identify the knowledge network domain which is defined by relative papers and their citations.

2). Data collection. We conducted a search on ISI Web of Science Core Citation Database. Retrieval strategy is “Topics= knowledge network AND Type= article AND Language= English”, with time span from 2007/01/01 to 2016/08/31. We removed duplicates by CiteSpace function. Finally we got 28580 bibliographic records.

3). Time slicing. We specify 1 year as the length of a single time slice.

4). Threshold selection. We selected top 50 references per time slice. In terms of threshold selection, CiteSpace allows threshold settings in three time partitions (before, during, and after) respectively at the layers of citation quantity, co-citation frequency, and co-citation factor, while other thresholds are determined by the linear interpolated values.

5). Pruning and merging. Minimize spinning tree is chosen for network pruning.

Data collection and sample

Scientific literature has both persistent and transient elements. The transient aspect of scientific literature can be characterized by corresponding thematic trends, while the persistent aspect can be characterized by salient conceptual structures (Fang, 2015). The analysis of thematic trends is based on the concept of burst detection (Chen and Song et al., 2008). We use the information contained in different fields for a standard scientific and a text-based analysis. CiteSpace can detect the emerging trends and tracking topics. We get the research field on knowledge network by visualizing mapping. CiteSpace was developed by Mr. Chen Chaomei (Cheung Kong Scholar Lecture Professor at Dalian University of Technology, and Chinese scholar at the Drexel University, USA) and his research team. It is a Java application used to identify and visualize the new trend and development in scientific researchers from scientific literatures and has become information visualization software that has great impacts in the information analysis field. Therefore, the CiteSpace is adopted to draw the Mapping of Science to disclose
the development of the knowledge network studies in China in visual mode and provide references for researchers in China, which is undoubtedly of great significance and value.

Secondly, we construe the model to propose an implementation process for literature-based domain knowledge network of a particular field, in which knowledge network is divided into structure network and relationship by choosing different information as the actor or link of the network. On the basis of the model, dimension-time, network type and level are designed to map into a cube to constitute the methodology structural framework for the purpose of creating and analyzing the domain knowledge network.

Results

Types of domain knowledge network based on Bibliometrics

The knowledge network can be divided into various types in different perspectives. It can be divided into macro knowledge network and micro knowledge network according to the level of view; it can be divided into explicit knowledge network and tacit knowledge network according to the explicit and implicit. It is meaningful to divide in accordance with the specific research object, because each domain of knowledge network is composed of node and link, different types of knowledge network can be obtained by selecting different nodes and links of knowledge. This paper argues that the knowledge network can be divided into two categories, the network of literature knowledge and the network of knowledge itself. The network of literature reflects the relationship between the existence of the association, inheritance and development, which can be constructed by the content and citation attributes. Knowledge itself reflects the cooperation, competition, etc. in different body of knowledge, which can be constructed by the attribution of author or organization, etc. in view of considering the specific application, knowledge has timeliness, and changed with the time changes. Therefore, the knowledge network is actually composed of nodes, association and time, which can form a cube of the knowledge network. (See figure 1)
Figure 1. The constructive model of knowledge network in different domain

We can get four network types from the perspective of network in Fig. 1, which includes SN network, co-word network, citation network and cooperation network; from the perspective of research level, it has macro, medium, micro network; from the perspective of time, it divided by year. SN network focuses on the similar network, which describes the relationship between two articles by keywords, for example, article A is a node in network, other article with the same keywords like article A will link to article A. these two articles are the similar research articles, which uses this method to build the network called SN network. Co-word network is the link between words when a group of keywords in the same article, which founded by the method of network known as the co-word network. Obviously, when the link is regarded as the relation between knowledge, knowledge network reflects the relation network between the knowledge itself. The relationship network of knowledge includes cooperation and citation. Cooperation between authors, institution reflects the relation network in knowledge citation, and knowledge transfer. Citation reflects the relationship between inheritance and development of predecessor’s knowledge, which explains the knowledge transfer from cited literature to citation. Thus, the types of cooperation networks and citation networks can be formed.

The realizing process of knowledge network in different domain based on literature

Knowledge is divided into tacit knowledge and explicit knowledge, which can be divided into two categories: literature knowledge and knowledge agent. Literature knowledge is a kind of explicit knowledge, while knowledge agent is the creator of literature knowledge and the inheritor of tacit knowledge. In view of the relationship of inheritance and development in domain knowledge, the paper establishes the model of knowledge network in different field form the perspective of content attribute, citation attribute. The structure and distribution of the domain knowledge network can be reflected by strength of these relationships. The different knowledge agent can produce
collaboration, learning, and competition between each other in the course of interaction with knowledge, which makes the knowledge flow and transfer. Therefore, it can build the knowledge agent networks by the link among the knowledge. The author attributes is as a unit of knowledge based on analysis of construction for knowledge network. The network of knowledge agent and knowledge structure are not two independent networks, and they are two kinds of static and dynamic forms of knowledge network in a certain field, which is one subject of two aspects. The network knowledge structure is the foundation of the network of knowledge agent, while the network of knowledge agent is the extension of the knowledge structure. The process of literature analysis to the realization of knowledge network is as showed in figure 2. From the figure, the model constructs “knowledge unit”, “extracting knowledge node”, “establish relation”, and “build visualizing network”; in the horizontal dimension, the model distinguish the characteristics of various knowledge networks and relationship between each other.

![Diagram of Knowledge Network](image)

**Figure 2. The realizing process of domain knowledge network in different domain**

In the knowledge network, the research of drawing tool is one of the main contents. It is very important to develop the visual software which is powerful, easy to use and display the image. The drawing of knowledge network cannot be separated from the use of a variety of visualization tools. Used more tools or software currently include CiteSpace, as well as social network analysis software UNCINET and Pajek from analysis of the literature. The Indiana University Library and Information Science Institute developed the large-scale network data analysis, modeling and visualization tools that are Network Workbench Tool and Sci 2 Tool. VOSviewer is developed by Holland scholar Van Eck and Waltman. Vanagepoint is mostly used by researchers. Table 1 describes the situation of these visualization tools. At present, more and more visualization tools used in the literature, which can be easily and intuitively displayed.
the research problems in front of scholars by using visual tools. This is one of the reasons for the increasing in the use of visual tools in the literature.

Table 1. The visualization tool of knowledge network

<table>
<thead>
<tr>
<th>software</th>
<th>open source</th>
<th>common method</th>
<th>development institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>citespace</td>
<td>Yes</td>
<td>burst detection, spatial analysis, construct network, time sequence analysis</td>
<td>Drexel University</td>
</tr>
<tr>
<td>Gephi</td>
<td>No</td>
<td>community detection, dynamic network analysis</td>
<td>Association for the Advancement of Artificial Intelligence</td>
</tr>
<tr>
<td>In-SPIRE</td>
<td>Yes</td>
<td>burst detection, construct network, time sequence analysis</td>
<td>Pacific Northwest National Laboratory</td>
</tr>
<tr>
<td>Network Workbench Tool</td>
<td>Yes</td>
<td>burst detection, construct network, time sequence analysis</td>
<td>Indiana University</td>
</tr>
<tr>
<td>Pajek</td>
<td>Yes</td>
<td>construct network, time sequence analysis</td>
<td>Ljubljana</td>
</tr>
<tr>
<td>Sci2 Tool</td>
<td>Yes</td>
<td>burst detection, spatial analysis, construct network, time sequence analysis</td>
<td>Indiana University</td>
</tr>
<tr>
<td>SciMAT</td>
<td>Yes</td>
<td>construct network, time sequence analysis, evaluation of performance and quality</td>
<td>Granada University</td>
</tr>
<tr>
<td>Ucinet</td>
<td>Yes</td>
<td>construct network</td>
<td>Analytic Technologies</td>
</tr>
<tr>
<td>VanagePiont</td>
<td>No</td>
<td>burst detection, spatial analysis, construct network, time sequence analysis</td>
<td>Search Technology</td>
</tr>
<tr>
<td>VOSViewer</td>
<td>No</td>
<td>community detection, construct network</td>
<td>Leiden University</td>
</tr>
</tbody>
</table>

Method of domain knowledge network based on literature

To draw the domain knowledge network based on literature is to study and solve the problems in the academic research, such as measuring the strength of collaboration between the author, etc. It is a leap in the process, and a scientific and constructive work from the visualization to the conclusion of research. The process mainly includes two kinds: one is to judge directly from the graph description, such as discovering the size of knowledge unit, strength and their relationship by the nodes, color, size, line of connection, shape of network, etc.; the other is judging by the measurement of relative index. In this paper, we consider that the knowledge network can be measured by 4 basic indicators, which include the breadth of knowledge network, the depth of knowledge network, the strength of collaboration, the relation of knowledge link and arcs. The breadth of knowledge network refers to the number of links in the knowledge network, which means other nodes are connected to the target node. The depth of knowledge network is the number of intermediate nodes that are contained between knowledge units or knowledge nodes, the more nodes, the greater relationship, which indicates whether the knowledge nodes have direct or indirect correlation. The strength of collaboration is closely related to the knowledge nodes that are direct relation. It is generally expressed by the distance of the knowledge node, the length of the connection, and the thickness.
of the connection. The relation of knowledge link and arcs indicates that the knowledge unit exists identity relation, equivalent relation, hierarchy, rank correlation, inclusion relation, causal relation, parataxis, etc. the arcs includes in degree and out degree, which indicates the direction of knowledge node.

The basic description of the structure for knowledge network mainly includes average path length, clustering coefficient, and connectivity. The average path length refers to the average value of the shortest path between any nodes of network, which can be used to measure the distance between nodes. The clustering coefficient of network can measure clustering feature, which is the average value of all clustering coefficient, and the clustering coefficient of nodes refers to the ratio of the number of triangles formed by the nodes and the number of elements. The out degree of node is one node points to the number of edges of the other nodes; in degree of node is the number of edges form other nodes pointing to slide of one node, and the connectivity is a sum of out degree and in degree. The proportion of papers that are directly judged by knowledge network is relatively more, which reflects the advantage of the method in intuition and visualization. At the same time, it also shows that some researchers have a lack of understanding for the mechanism of visualization software.

Panorama of domain knowledge network

The intellectual landscape of the research field can be represented by CiteSpace, which can draw a network of cited references, categories and co-occurring keywords. Our study focuses on co-cited references network, co-occurring keywords and categories. Figure 3 presents the top cited literatures associated with the term “knowledge network” for 2007 to 2016. The graph consists of 28580 records, which have the highest betweenness centrality in the network, suggesting that each records as an essential resource in knowledge network. All the articles use knowledge network as a key enabling technology for pursuing research questions. As showed in figure 1, modularity Q 0.819 and mean silhouette 0.3143 indicates a good inter-cluster connections network and considerable partition of the network. In the map, there are several purple round nodes, which reveal that these are developed area of the knowledge network. There are mainly two clusters in the map, which represents knowledge subject relation network and knowledge structure network.
Figure 3. Panorama of knowledge network based on literature from 2007 to 2016

Which disciplines are involved in knowledge network? Each article indexed by the Web of Science is assigned one or more subject categories. Figure 4 shows a network of such subject categories by Pathfinder network analyzing, which indicates the most salient connections. The most common category is management, library and information science, education, computer science artificial intelligence, computer science information systems, business, mathematical computational biology, etc. The most common category is Management, which has the largest article, followed by Computer science and Business & Economics, Engineering. From the figure 4, the method of knowledge network is used in more and more disciplines.
Keywords are the refined research content in the article, and the high frequency words appear in the article and the noun phrase extracted from each article can be regarded as a research hotspot in the field. With the help of visualization software, CiteSpace, to draw the clustering map, to detect the hotspot in the field of research. We detect the high frequency words in the field of knowledge network by drawing the knowledge network. We set the appropriate parameters in the CiteSpace and select the pathfinder algorithm to draw the map, which is shown as in figure 5. Pathfinder relies on the so-called triangle inequality to eliminate redundant or counter-intuitive links, which are a complex network that represents proximity data to a much more concise and meaningful network-only the most important links in the network (Chen, 1999). The circular nodes in the map represent the key words, the size of node and its tag is proportional to the frequency of the words, and the larger nodes can be regarded as the main research topics in the field of knowledge network from 2007 to 2016. From the figure 5, we can get the larger frequency keywords as showed in table 2.

World frequency analysis is the Bibliometric methods to determine the research hotspot and developing trend in the research area. It uses keywords or subjects that appear on the frequencies in the area to explain or express the core content of literature. Due to the keywords or subject are the core content of the article, or the concentrated and refined content in the literature. Therefore, it can reflect the hotspot of the research, if the keywords or subject appear repeatedly in the literature. Table 2 is the main high-frequency keywords and high-betweenness Centrality that reflect the main research area of knowledge network by visualization software, CiteSpace. The high-betweenness centrality of a node is a graph-theoretical property that quantifies the importance of the node’s position in a network (Freeman, 1979). The high-betweenness centrality keywords tend to be found in paths connecting different clusters. This feature has been
used in community-finding algorithms to identify and separate clusters (Girvan and Newman, 2002).

Table 2. High-frequency keywords and high-betweenness keywords

<table>
<thead>
<tr>
<th>Rankings</th>
<th>High-frequency Keywords</th>
<th>Frequency</th>
<th>High-betweenness Keywords</th>
<th>Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>system</td>
<td>692</td>
<td>gene expression</td>
<td>0.16</td>
</tr>
<tr>
<td>2</td>
<td>model</td>
<td>649</td>
<td>evolution</td>
<td>0.16</td>
</tr>
<tr>
<td>3</td>
<td>innovation</td>
<td>566</td>
<td>innovation</td>
<td>0.15</td>
</tr>
<tr>
<td>4</td>
<td>management</td>
<td>427</td>
<td>architecture</td>
<td>0.14</td>
</tr>
<tr>
<td>5</td>
<td>neural network</td>
<td>407</td>
<td>management</td>
<td>0.13</td>
</tr>
<tr>
<td>6</td>
<td>social network</td>
<td>349</td>
<td>identification</td>
<td>0.11</td>
</tr>
<tr>
<td>7</td>
<td>organization</td>
<td>338</td>
<td>model</td>
<td>0.08</td>
</tr>
<tr>
<td>8</td>
<td>algorithm</td>
<td>289</td>
<td>neural network</td>
<td>0.08</td>
</tr>
<tr>
<td>9</td>
<td>information</td>
<td>266</td>
<td>social network</td>
<td>0.08</td>
</tr>
<tr>
<td>10</td>
<td>technology</td>
<td>249</td>
<td>technology</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Theoretical basis of knowledge network method based on Bibliometric

The method of knowledge network based on Bibliometric is widely used in current academic research, which has a very complicated reason.

(1) The growing importance of knowledge network in the academic field

From the perspective of academic research to analyze, the figures used mainly after geometric interpretation was founded by Descartes and Fermat. The academic people have further understanding for the advantage and characteristics of graph with the development of scientific technology. Such as, figure can break the limitations of time and space to make some of invisible phenomena understand easily. It can help people to discover new information by analyzing the picture, and it can easily establish the relation between experience and theory, which makes the mathematical analysis and scientific computation to carry out smoothly. Knowledge mapping is invariance in the content, in addition to mobility, intuitive and visualization. It can be kept relatively stable whether it is the summary of scientific phenomena or during the process of academic exchange communication. The knowledge mapping is the powerful tool for the scientists to use in the academic debate. The science begins when scientists separate themselves from nature into inscription of the world. Therefore, knowledge mapping plays a key role in integration of facts and theories. The knowledge mapping often shows the hidden rules because it is easy to be combined with geometry, so it has a strong ability to inspire and analyze. The application and attention in the academic field are increasing. Some academic field considers the knowledge mapping as one of the important signs in surveying the development of discipline.

(2) The development of social sciences in social network analysis method

The study of social network in various social and organizational fields has undergone extensive development over last three decades (Zaheer and G Zübüyük et al., 2010). Social network analysis is the process of investigating social structures through
the use of network and graph theories (Ozyar and Gurdalli, 2002). The social network concentrated on both personal and organizational networks with the discussing on interdisciplinary nature of network. It characterizes networked structures in terms of nodes and ties, edges, or links that connected them, which has an earlier conceptual background. It was in the 1960s that researchers from the University of Harvard included mathematical developments in the study of social relations. Social network analysis covers a lot of methods that reveal the structures of the relational data that are in social practice. With the development of social network analysis, it has formed a series of terms and concepts, as a very useful research method and thinking mode. It has already went beyond the scope of the field of sociology, which is widely used in various fields, including citation network, co-occurrence keywords, co-author citation within the field of Library and Information Science.

(3) The research and development of the network in the field of Scientometrics

The research of knowledge network begins with relationship network of literature in the field of Scientometrics. The representative literatures are as following, in 1955, a new Dimension in Documentation through Association of Ideas wrote by Eugene Garfield who propose a bibliographic system for science literature that can eliminate the uncritical citation of fraudulent, incomplete, or obsolete data by making it possible for the conscientious scholar to be aware of criticisms of earlier papers (Garfield, 2006). In 1965, Networks of Scientific Papers was wrote by Price who described in the broadest outline the nature of the total world network of scientific papers (Price, 1965). In 1973, Co-citation in the Scientific Library: A New Measure of the Relationship between Two Documents was wrote by Henry Small who defined the new kind of “coupling” and to distinguish it from bibliographic coupling, using an actual example to agree generally with patterns of direct citation (Small, 1973). In 1998, Perfetti wrote one article about co-word, and he did a detailed description, after decades of development, co-word analysis is becoming more mature; people can find affinities with researching object by co-word analysis; it can also reveal implicit or potentially useful knowledge to discover the relationship of structure and subject (Perfetti, 1998). In 2003, mapping knowledge domains was presented at a workshop of National Academy of Sciences. Various tools of mapping knowledge domains are also increasing. Mapping knowledge domains was promoted by the famous American expert, Chen Chaomei who is from Drexel University. He developed the software, CiteSpace, which is to become the most popular visualization software in academic research. “Mapping Scientific Frontiers: The Quest for Knowledge Visualization” and “Information Visualization: Beyond the Horizon” are the classics in the field of information visualization or knowledge visualization (Chaomei Chen, 2003; Chen, 2006).

(4) The development of complex network theory in the field of physics and systems science

Complex network is a graph with non-trivial topological features that do not occur in simple networks such as lattices or random graphs but often occur in graphs modelling of real systems. It is a hot topic in physics and systems science, which has become a new interdisciplinary in 21st, and it concerns with the general characteristics and universal methods of complex networks to promote the understanding of natural and artificial complex networks. People have been filled with a variety of complex networks in
nowadays. The network has become an indispensable part of life in modern times, such as the World Wide Web is composed to web pages one by one; the power network by the power station; food chain and ecological network are composed by species; social network is composed of family, neighbors, friends etc. Therefore, the world was filled with various complex networks, whether it is tangible or intangible, animate or inanimate, artificial or natural, even as the key words within the literature, which seems to not a network of things that can be studied from the perspective of complex networks.

(5) The rapid development of data visualization technology in the field of computer science

Data visualization technology is to use computer to deal with mass of data to the graphic or image displayed on the screen by computer enormous processing capacity and graphic processing technology. It involves computer graphics, image processing, computer aided design, computer vision and human-computer interaction, etc. the concept of data visualization is derived from visualization in scientific computing, but with the development of computer technology, the concept of data visualization has been greatly expanded. It includes not only the visualization in scientific computing, but also the visualization of engineering data and measured data, as well as the development of information visualization in recent years. Visualization of scientific computing is an interdisciplinary branch of science, which considered a subset of computer graphics, a branch of computer science (Lopes and M et al., 2010). The purpose of visualization of scientific computing to enable scientists to understand, illustrate, and glean insight from the data. Therefore, visualization in scientific computing refers to the visualization of spatial data sites, while the information visualization refers to non-spatial data. The application of data visualization has been very extensive, covering various fields, such as natural science, engineering technology, finance, communication and commerce. Visualization can be used to discover the complex relationships among the knowledge units as a powerful tool for knowledge discovery, and it has been paid attention by the researchers of the library and information science.

Discussion and conclusion

The domain knowledge network includes the structural network of domain knowledge network itself, and the relation network among knowledge agent. The paper put forward the realizing process of domain knowledge network from knowledge and knowledge agent and elaborated the construction of knowledge network model form different domains. We constructed the domain knowledge structure and relation network from time factor, network dimension and research domain based on Bibliometrics to reflect the domain knowledge network further.

The literature is a carrier that contains the information of scientific research achievements and scientific research activities. It can help us to solve many academic problems by domain knowledge network method based on literature. The domain knowledge network method has a special value to researchers. It has been widely used in many fields of discipline, such as scientific research, library and information science, management science, biology and medical informatics, as well as education, sports, economics and engineering technology, etc. Therefore, it is conductive to the further understanding of the academic field and the correct use of the method from the
perspective of methodology of theory to sort, profound consideration, review, and prospect.

Limitation

As a research method, there are some deficiencies in the academic field by using domain knowledge network method based on literature, mainly in the following aspects: ① Currently, studies on the knowledge network focus more on the visual form of the knowledge network structure and description on the relationship, but less on the dynamic process and evolutionary mechanism of the knowledge development. ② It is not enough to research in formation process of the hotspot and innovation trend for the evolution of knowledge network, and the prediction is not strong. ③ The strategy is inappropriate for data downloading, which cannot reach the integrity and accuracy of the data. ④ Each visual software has advantage and disadvantage. It should choose the appropriate tools for different analysis. There is no perfect visualization tool.

Acknowledgments

This paper is supported by National Social Science Foundation in China (Grant No.16BTQ055)

References


Abstract
In this contribution a bibliometric analysis of the computer science theses defended in Spain is carried out based on bibliographic networks and communities detection. Data was retrieved from Teseo database. Then, a filtering process was applied in order to get only theses data in which doctoral programs title match with keywords related with Computer Science. Then, we build a network with theses defense committee members from data extracted and we apply fundamentals of social network analysis and community detection techniques to analyze data and uncover novel characteristics from network structure.

Conference Topic
Country-level studies
Mapping and visualization
Social network analysis
Co-occurrence analysis

Introduction
A doctoral thesis is a research job carried out by a doctoral student to obtain a postgraduate level with a higher academic rank. This consists in carrying out a rigorous research process about a chosen topic, which should have clear objectives, hypotheses and a research plan. Also it must have a detailed report of the methodology used and the results achieved, with the aim of generating significant and original contributions for a knowledge area.

Once the thesis has been completed and accepted, the student proceeds to defend his/her thesis before a committee that can be made up of three or five members, of which at least one third or two fifths (in the case of five members) must belong to the university in which the postgraduate studies have been carried out. In addition, the members of the committee can be foreign, depending on the area of knowledge in which the thesis has been developed. After the thesis defense, the student will therefore sign a document, whereby he or she, the author, permits the consultation and the lending of the thesis.

Once a thesis has been defended, it is stored in a web repository of public access, of the Secretariat of the Council of Universities (Teseo). This website allows to find the doctoral dissertations defended in Spain according to the official register of the Ministry of Education. Therefore, anyone can access this information. For instance, the author, supervisors, defense committee, title, abstract, keywords, among others.

Therefore, this information could be used to carry out a science mapping analysis (Cobo, López-Herrera, Herrera-Viedma, & Herrera, 2011b) based on bibliographic networks (Batagelj & Cerišek, 2013) in order to uncover the social (Newman, 2004) (Peters & Raan, 1991) (Glänzel, 2001), intellectual (Small, 1973) (Kessler, 1963) and conceptual (Callon, Courtial, Turner, & Bauin, 1983) aspects (Cobo, López-Herrera, Herrera-Viedma, & Herrera, 2011a).
Particularly, the relationships among the committee members could be modeled as a social bibliographic network, where the nodes will be the members, and the edges will represent members co-occurrence and this could be used to analyze patterns of collaboration of members. In Bibliometrics, Collaboration Networks could be analyzed using indicators to measure and extract the structure inherent to a set of publications; but also these could be examined as social networks to explore interactions between actors who may be researchers, groups, institutions, etc. that collaborate in the development of scientific papers on research topics (Newman, 2004). The study of these networks is done in the beginning to study the behavior of those involved and also to find patterns of collaboration. Newman et al. conducts a study of the collaborative networks between scientists in the field of physics, biomedical research, and computer science (Newman, 2001). These have all the characteristics of small worlds networks: small distance between nodes, a high clustering coefficient and the degree of distribution seems to follow a power law (Barabási et al., 2002).

So, the main aim of this contribution is to perform a science mapping analysis based on social bibliographic networks, using as data the relationships of the Spanish theses defense committee members gathered from Teseo. Particularly, our objectives are: propose Teseo as new bibliographic source for bibliometric analysis, build a theses committee bibliographic networks, perform a social network analysis, and detect the hidden social groups by means of community detection.

The rest of the paper is organized as follows: Section Methodology, shows the steps followed to carry out the analysis. Section Results and statistics, shows the output of applying the steps stablished in the previous stage. And finally, some conclusions are drawn.

**Methodology**

We have followed a set of procedures to achieve a science mapping analysis of doctoral theses data from Spain. Specifically, the object of study will be theses defense committee members. For this, the methodology shown in the Figure 1 has been established and is described below step by step.

**Data Collection**

For this study we have used data from Teseo database. Teseo is a huge repository which contains doctoral theses of all universities from Spain. It is fed on the information provided by the doctoral commissions of different universities and is regulated by the University Coordination Council. This contains information of theses defended since the year 1976 and this information is released for searches at URL: https://www.educacion.gob.es/teseo/irGestionarConsulta.do.
Thus, Teseo provides information like thesis author’s name, thesis title, thesis directors, court before which the thesis was defended, the program on the thesis was developed, university, thesis reading date, a summary of thesis topic, among other relevant data. To be able to manipulate data of Teseo web page, first, we have had to download it and for that we have constructed a small code including the combined use of Python language and the BeautifullSoup library\(^1\) as tools to extract and store great volume of structured data in web page. In this script we have considered that the URL with the information of each thesis is formed in the same way, except for a reference that is a sequential numerical code that differentiates each thesis. In such a way that the information of the same thesis is repeated every 3 reference codes. Therefore, this code runs through each URL and stores in a structure all the data of the website that are later written in a CSV file. The following Teseo information has been downloaded:

- Reference
- Title
- Author
- College
- Department
- Date
- Program
- Director
- Doctoral court
- Descriptors

Collected data has been saved in a csv file and then stored into a relational database.

**Data Storage**

We have used MySQL to store data from CSV file for future analysis but mainly to review if the data is correct, and looking for missing values, or inconsistent records. Initially 237187 theses records have been downloaded from Teseo web page, but, to do this analysis more specific, we have decided to manually filter data to focus on the study of the theses whose doctoral programs are related to the area of computer science. Thus, in Table 1 the selected keywords used to filter the theses are shown.

<table>
<thead>
<tr>
<th><strong>Keywords</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPUT*</td>
<td></td>
</tr>
<tr>
<td>INFORM<em>T</em></td>
<td></td>
</tr>
<tr>
<td>SOFTWARE</td>
<td></td>
</tr>
<tr>
<td>ORDENADORE?</td>
<td></td>
</tr>
<tr>
<td>ARTIFICIAL</td>
<td></td>
</tr>
<tr>
<td>INTELIGEN*</td>
<td></td>
</tr>
<tr>
<td>TECNOLOG* INFORMA*</td>
<td></td>
</tr>
<tr>
<td>CIBERN*TICA</td>
<td></td>
</tr>
</tbody>
</table>

The distribution of theses per year is shown in Figure 2 where we can see the distributions of number of theses per year for all Teseo data and in Figure 3 is shown the distribution of number of theses about Computer Science programs per year.

---

\(^1\) [https://www.crummy.com/software/BeautifulSoup/bs4/doc/](https://www.crummy.com/software/BeautifulSoup/bs4/doc/)
After filtering process, we obtain 6528 theses with doctoral programs related to computer science and also have the names of defense committee members.

**Data Pre-processing**

A dataset composed by 6528 records (target data) have been pre-processed using RStudio tools to do different steps, which as described as follows:
As we intend to study the relationships between defense committee members, we have first had to process the Tribunal column from dataset because it can contain three or more member’s names of defense committee in a single row. Then it has been divided, so that each row contains only one member’s name to do a granular data analysis.

After we have implemented an R function to convert data, we have get a new dataset with 29937 rows instead of 6528. With this new dataset we have proceeded to build the graph and its corresponding adjacency list to relate the member’s names of defense committee who have worked together. For this we have used Igraph\(^2\) package in R. Igraph is a collection of network analysis tools with the emphasis on efficiency, portability and ease of use. It is open source and free and can be programmed in R, Python and C/C++. It provides routines and functions to create and manipulate graphs easily.

So, the adjacency list will be used to create networks that will show us the interaction of different member’s names of defense committee.

As example, let suppose a set of six theses. Each thesis has a defense committee as it is shown in Table 2. Thus, the network built from this example corpus are shown in Figure 4. We should point out that an arc between two nodes represents a relation between these committee members. Also, the thickness of the arcs represents the raw co-occurrence.

<table>
<thead>
<tr>
<th>Thesis</th>
<th>Defense committee members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thesis 1</td>
<td>d1,d2,d3</td>
</tr>
<tr>
<td>Thesis 2</td>
<td>d4,d5</td>
</tr>
<tr>
<td>Thesis 3</td>
<td>d4,d3,d6,d7,d8</td>
</tr>
<tr>
<td>Thesis 4</td>
<td>d2,d8,d3,d4,d5</td>
</tr>
<tr>
<td>Thesis 5</td>
<td>d2,d3</td>
</tr>
<tr>
<td>Thesis 6</td>
<td>d2,d3</td>
</tr>
</tbody>
</table>

**Figure 4 Network with different sizes and different weights.**

Therefore the network will have some properties established in the scope of this analysis. On the one hand, nodes will have two properties: name and frequency. The name property is used to record the defense committee member’s name who is represented by the node and the frequency property is an integer that represents the number of times a name

\(^2\) http://igraph.org/r/
appears in the graph, this means, if a defense committee member appears 100 times then this will have a frequency equal to 100. This property has been established within the framework of this methodology, to assign a level of importance to defense committee members, therefore, those who have higher frequency values will be considered more important than those with low frequency values.

On the other hand, edges or relationships also have properties. In this case, it has been defined weight property, which represents the number of times that is found an interaction between two members.

**Graph Construction**

The adjacency list gotten in the previous stage has been used to build an undirected graph using Igraph package in RStudio with all the properties named above.

We have decided used Igraph because is a very complete library that allows the construction of graphs from structures such as data.frames in R. In addition, it allows to apply a number of methods, routines and functions to manipulate graphs facilitating the analysis of networks.

One problem with large-scale networks is that they are difficult to analyze with the naked eye. To avoid having a graph of a very dense network where we cannot easily appreciate the relationships between the nodes, we have proceeded to prune the network excluding those nodes that have relationships with a weight less than 2.

**Graph Analysis**

With the network built above we have used the Fast Greedy (Clauset, Newman, & Moore, 2004) algorithm to detect communities with the aim of discovering the structure of the network. We have applied the Igraph function that allows to assign a group to each node of the network, using the algorithm.

For graph visualization, VisNetwork library (also available in R) has been used because this library draws networks with a dynamic aspect, allowing to zoom on the network, to add tooltips, to assign different forms for the nodes, select nodes neighbors, etc.

In addition, centrality measures of nodes have been calculated to detect important nodes of the network (Freeman, 1978). We have used Igraph functions to get Degree, Closeness, Eigenvector and Betweenness from network nodes.

Referring to the concepts of centrality, it has been established some attributes:

- **Best connected member** that represents one or more theses defense committee members with the highest Degree Centrality value in the network. This means that a member has a privileged position within the network because he/she has been on a committee with more variety of people than the others. That is, it has a greater amount of connections with other different individuals. His/her importance is based on that he/she has an advantage because it is supposed to interact with more people. He/she is more independent than others and more powerful.

- **Member with the most closeness** that represents one or more members with the highest Closeness Centrality value in the network. This means the importance of the member is defined by the ties of his/her neighbors. If his/her neighbors have more connected nodes, he/she is more important.

- **The most intermediary member** that represents one or more members with the highest Betweenness Centrality value in the network. This means that the member is more accessible by others. His/her importance consists in that while more nodes need to go through him to access others he/she will have more interactions.

- **The most influential member** that represents one or more members with the highest value in the Eigen vector calculated over the network. This uncover the importance of a
members in terms of the overall structure of the network. It says a member is important if he/she is linked to more important members. Therefore, his/her power in the network depends on his/her neighbors too. A committee member with high degree connected to members with high degree too is more influential.

Then, we have used the communities’ sizes to get a subgraph with two most representative communities of the network. From this we can identify who are the defense committee members which are related to more of the others.

**Results and statistics**

Following all the steps described in the methodology we have gotten an undirected graph with 11963 nodes and 48980 edges. It is a very dense network, in which the main component is formed by 10647 nodes, while the rest of the components are formed between 3 and 20 nodes. Therefore, we have had to prune the graph as mentioned in the methodology section, excluding those members who are considered less relevant. We have gotten a new network with 465 nodes and 328 edges as can be seen in Figure 5 where the frequency of nodes is reflected in the size of these and the weight of the relationships is reflected in the width of the edges; however, it is difficult to understand just by seeing this because there is a lot of data in graph.

![Figure 5 Pruned Graph of Theses Defense Committee Members.](image)

On this network we have used the centrality measures to obtain the most important nodes, the following results are shown in Table 3 for the whole network and in Table 4 we can see the results for the pruned graph.

As in this case most of the times the measures point to the same node, placing this as the best node located in the network. In Figure 6 the pruned network is shown according to different measures. Thus, in Figure 6.a the sizes of the nodes is proportional to the degree, in Figure 6.b, the size of the nodes is proportional to the closeness centrality, In Figure 6.c the size of the
nodes is to betweenness, and finally, in Figure 6.c the size of the nodes is proportional to the eigenvector centrality.

Table 3 Centrality measures results for the whole network.

<table>
<thead>
<tr>
<th>Best connected member</th>
<th>Member with the most closeness</th>
<th>The most intermediary member</th>
<th>The most influential member</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toro Bonilla, Miguel</td>
<td>Herrera Triguero, Francisco</td>
<td>Toro Bonilla, Miguel</td>
<td>Toro Bonilla, Miguel</td>
</tr>
<tr>
<td>Piattini Velthuis, Mario</td>
<td>Toro Bonilla, Miguel</td>
<td>Piattini Velthuis, Mario</td>
<td>Piattini Velthuis, Mario</td>
</tr>
<tr>
<td>Pavón Mestras, Juan</td>
<td>Pavón Mestras, Juan</td>
<td>Herrera Triguero, Francisco</td>
<td>Herrera Triguero, Francisco</td>
</tr>
<tr>
<td>Herrera Triguero, Francisco</td>
<td>Troya Linero, José</td>
<td>Pavón Mestras, Juan</td>
<td>Pavón Mestras, Juan</td>
</tr>
<tr>
<td>Tirado Fernández, Francisco</td>
<td>Corchado Rodríguez, Juan Manuel</td>
<td>Larrañaga Mugica, Pedro</td>
<td>Larrañaga Mugica, Pedro</td>
</tr>
</tbody>
</table>

Table 4 Centrality measures results for the pruned network.

<table>
<thead>
<tr>
<th>Best connected member</th>
<th>Member with the most closeness</th>
<th>The most intermediary member</th>
<th>The most influential member</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muñoz Pérez, José</td>
<td>Garcia Fernández, Inmaculada</td>
<td>Muñoz Pérez, José</td>
<td>Hernández Bude, Porfidio</td>
</tr>
<tr>
<td>Viñals Yufera, Victor</td>
<td>Díaz Bruguera, Javier</td>
<td>Díaz Bruguera, Javier</td>
<td>Díaz Bruguera, Javier</td>
</tr>
<tr>
<td>Rodríguez León, Casiano</td>
<td>Muñoz Pérez, José</td>
<td>García Fernández, Inmaculada</td>
<td>García Fernández, Inmaculada</td>
</tr>
<tr>
<td>Hermida Correa, Román</td>
<td>Hernández Bude, Porfidio</td>
<td>Rodríguez León, Casiano</td>
<td>Rodríguez León, Casiano</td>
</tr>
<tr>
<td>Castro Sánchez, Jose Jesús</td>
<td>Luque Fadon Emilio</td>
<td>Luque Fadon, Emilio</td>
<td>Luque Fadon, Emilio</td>
</tr>
</tbody>
</table>
At granular level, we have chosen ten people who have higher frequency values. This means that they appear more often in the full network. In Table 5 is shown the list of members who have been part of the defense committee more times. In our analysis we have detected that those with higher frequency values do not necessarily coincide more times with the same people on the committee, at least in this area of knowledge.

Table 5 Frequency of occurrence of individuals.

<table>
<thead>
<tr>
<th>Theses defense committee member</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>96</td>
<td>Toro Bonilla Miguel</td>
</tr>
<tr>
<td>90</td>
<td>Tirado Fernández Francisco</td>
</tr>
<tr>
<td>87</td>
<td>Piattini Velthuis Mario</td>
</tr>
<tr>
<td>62</td>
<td>Pazos Sierra Juan</td>
</tr>
<tr>
<td>61</td>
<td>López Zapata Emilio</td>
</tr>
<tr>
<td>61</td>
<td>Herrera Triguero Francisco</td>
</tr>
<tr>
<td>61</td>
<td>Pavon Mestras Juan</td>
</tr>
<tr>
<td>59</td>
<td>García Fernandez Inmaculada</td>
</tr>
<tr>
<td>54</td>
<td>Luque Fadon Emilio</td>
</tr>
<tr>
<td>54</td>
<td>Ayguadé Parra Eduard</td>
</tr>
</tbody>
</table>
On the other hand, we have applied Fast Greedy algorithm for community detection because this offers a global optimal solution in a reasonable time. Greedy strategies are based on recursive tasks where the solution of a problem depends on solutions of smaller instances of the problem. We have used the pruned network to make visualization and understanding of results easier and we have discovered 107 communities. In Figure 7 is shown how the communities are formed where each community is represented by a color.

![Figure 7 Communities detected with Fast Greedy algorithm.](image)

We have also obtained the subgraph with the two most representative communities. With this we can conclude that there are a large group of people who have been members of the Committee who are frequently related. In Figure 8 we can see the biggest community from graph detected with Fast Greedy algorithm. This have 28 nodes and 35 edges. Meanwhile, in Figure 9 it is shown the subgraph of the second biggest community, which has 25 nodes and 29 edges.
Conclusion
In this contribution, we have made a bibliometric study about data from Spanish doctoral theses in Computer Science area. We have extracted data from Teseo database website with the aim of uncover some novel information. We have followed a set of steps to find communities inside the network of Theses defense committee members, and finally we have uncover some interesting results and statistic.
We have detected that data stored in Teseo database has some flaws, such as it is not normalized, the names of the programs do not usually coincide with each other even though they are the same, besides the names of people who are in this database are usually not well written what incurs in our analysis can find two or more nodes that actually represent the same person. For example, we have found DÍAZ BRUGERA JAVIER and DÍAZ BRUGUERA JAVIER which obviously refers to the same person but his/her name is misspelled. There are also incomplete data and fields entered in the wrong columns.

As future work we propose to perform a preprocessing step on the raw data in order to normalize this, as well as to include in the analysis other types of social relations such as the relationships between theses directors and members of the defense committee and relationships between theses directors and keywords where we will surely find interesting patterns of collaboration.

Acknowledgments
The authors would like to acknowledge FEDER funds under grants TIN2013-40658-P and TIN2016-75850-R, and also the financial support from the University of Cádiz Project PR2016-067.

Bibliography


Publication patterns in the social sciences and humanities in Flanders and Poland

Emanuel Kulczycki\textsuperscript{1}, Tim C.E. Engels\textsuperscript{2} and Robert Nowotniak\textsuperscript{3}

\textsuperscript{1}emek@amu.edu.pl
Adam Mickiewicz University in Poznań, Faculty of Social Sciences, Scholarly Communication Research Group (Poland)

\textsuperscript{2}tim.engels@uantwerpen.be
University of Antwerp, Faculty of Social Sciences, Centre for R&D Monitoring (Belgium)

\textsuperscript{3}rnowotniak@kis.p.lodz.pl
Lodz University of Technology, Faculty of Electrical, Electronic, Computer, and Control Engineering, Institute of Applied Computer Science (Poland)

Abstract
This paper investigates internationalization patterns in the language and type of social sciences and humanities publications in non-English speaking countries. This research aims to demonstrate that such patterns are related not only to discipline but also to each country’s cultural and historic heritage. We used data from Flemish and Polish databases collected between 2009 and 2014. In Flanders, on the one hand, we found that changes in the use of languages and publication types were moderate and occurred gradually over several years. In Poland, on the other hand, we found significant shifts in the use of certain publication types, sometimes from year to year. Examining the social sciences and humanities literature both as a whole and broken down by discipline, we observed similar variability over time in the proportion of work published in English and in article form. However, we found remarkable differences between Flanders and Poland regarding the most commonly used languages and publication types. Overall, we found few similarities between Flemish and Polish social sciences and humanities publication patterns.

Conference Topic
Country-level studies, Science policy and research assessment

Introduction
This study aims to advance the knowledge regarding social sciences and humanities (SSH) publication patterns in Europe. National studies on SSH research outputs in Finland (Puuska 2014), Flanders (Engels, Ossenblok, & Spruyt, 2012; Verleysen, Ghesquiere, & Engels), and Norway (Sivertsen, 2016) have reported stable patterns in terms of publication type, though in terms of publication language, academic work is gradually leaning toward greater English use. Furthermore, Sivertsen (2016) suggests that though publication patterns across countries are similar within SSH disciplines, these patterns differ between SSH disciplines. However, these studies have focused on Western and Northern European countries, rather than on Central and Eastern European countries, which have undergone various academic transformations over the past three decades, following the breakdown of Communist regimes (Kozak, Bornmann, & Leydesdorff, 2014; Kwiek, 2014). Thus, our contribution here is to demonstrate that in non-English speaking countries, internationalization patterns in the language and type of SSH publications are related not only to discipline but also to each country’s specific cultural and historic heritage. We analyze all scholarly SSH publications written by academics affiliated with Flemish and Polish institutions between 2009 and 2014. Though the presence of SSH publications in
databases like Scopus and Web of Science (WoS) has grown—and though, as a result, various research evaluation systems have come to use this presence as a criterion of productivity—these citation databases have limited coverage, particularly of scholarly publications from non-English speaking countries (Sivertsen 2014). Therefore, to develop a broader picture of publication patterns in Europe, we examine the Flemish and Polish academic databases, which offer a complete representation of SSH publications within their respective countries. We begin by describing the structure of the Flemish and Polish academic databases and defining the various publication types and inclusion criteria. We then briefly explain how the databases classify fields and disciplines, as well as how comparisons between SSH publication patterns in Flanders and Poland are made possible using both databases. Next, we describe the data and methods, as well as our findings regarding SSH publication patterns related to language and type. We conclude by discussing and interpreting the results in a broader context.

**Design and Structure of the Flemish and the Polish Databases**

Over the past two decades, several European countries have developed research evaluation systems for assessing national research productivity based on the number of published works. For instance, in the U.K., disciplinary panels evaluate the quality of publications, whereas Norway, Denmark, Finland, Flanders, and Poland have developed full-coverage bibliographic databases to underpin local performance-based research funding systems (Hicks 2012). The most well-known of these databases is the Current Research Information System in Norway database, or the CRIStin database, which collects bibliographic information from all fields and all higher education institutions. In 2008, the Flemish government established the Flemish Academic Bibliographic Database for the Social Sciences and Humanities (VABB–SHW, www.ecoom.be/nl/vabb), which collects all bibliographic references to published SSH research outputs by scholars affiliated with universities in Flanders, the northern Dutch-speaking region of Belgium. Similarly, the Polish government developed the Polish Scholarly Bibliography (PBN, www.pbn-ms opi.org.pl) in 2011 to underline its performance-based research funding system. Like CRIStin, the PBN collects bibliographic information from all science fields and all higher education institutions.

*The Flemish Academic Bibliographic Database for the Social Sciences and Humanities (VABB–SHW)*

Earlier studies have described the VABB–SHW’s methodology for collecting bibliographic data (Ossenblok & Engels, 2015; Verleysen et al., 2014). Five Flemish universities annually provide bibliographic information for publications from the previous two years. The interuniversity Centre for Research and Development Monitoring (ECOOM) serves as the database coordinator and technical operator. Moreover, the Flemish government has established an authoritative panel comprising 18 professors affiliated with Flemish universities, and this panel evaluates whether journal and book publishers fulfill the VABB–SHW’s inclusion criteria.

*The Polish Scholarly Bibliography (PBN)*

No English publication has described the PBN’s design and structure. However, the database’s records are used for Poland’s performance-based research funding system, and so the inclusion criteria and publication forms have been described in previous studies (Kulczycki, 2017; Kulczycki, Korzeń, & Korytkowski, 2017).

In Poland, all scientific units, including faculties and basic and applied research institutions, among others, must submit to the PBN all bibliographic information for their affiliated scholarly publications. The database is updated daily, and the data is verified by PBN editors.
and higher education administrators. Each scientific unit must submit its data at least once every six months. However, missing data can be added until the next cycle of scientific unit evaluation, which is conducted every four years.

Publication Forms and Inclusion Criteria

The VABB–SHW classifies each publication into one of five publication types: (1) journal articles, (2) monographs, (3) edited books, (4) book articles or chapters, and (5) proceedings papers that are published independently of special journal issues and edited books. The PBN assigns publications to one of four publication types: (1) journal articles, (2) monographs, (3) edited books, and (4) book chapters. It is important to note that in Poland, proceedings papers are classified as chapters in edited books. Hence, by merging the Flemish database’s “book articles and chapters” and “proceedings papers” classifications, these two classification systems can be compared.

Table 1 presents the inclusion criteria for each publication type in the VABB–SHW and the PBN. All publications recorded in these databases are peer-reviewed papers that contribute to the development of new insights or applications resulting from these insights (the Flemish criterion) and present an original research problem (the Polish criterion). Universities can also submit records with incomplete bibliographic metadata. In Flanders, such publications cannot be approved by the authoritative panel or contribute toward the performance-based research funding system, but in Poland, such publications may or may not be approved by the expert panels during the evaluation of scientific units. In our study, we analyze all publication types, whether approved or unapproved, because the PBN’s design and structure do not allow us to make this distinction.

Field and Discipline Classifications

In Flanders, each SSH publication is assigned to at least one of sixteen SSH disciplines (see Table 2). Moreover, publications are classified into three general categories based on the authors’ research affiliations with an SSH unit (i.e., a research group, research center, institute, or department).

In Poland, the science field classification includes eight areas, 22 fields, and 102 disciplines. For research evaluation purposes, each publication is first assigned a classification based on the authors’ affiliations with a scientific unit, and then the scientific unit is given two classifications:

1. Scientific unit types: (a) faculty, (b) applied research institute, (c) basic research institute at the Polish Academy of Sciences, and (d) other.
2. Science fields: (a) social sciences and humanities, (b) life sciences, (c) sciences and engineering, and (d) arts sciences and artistic production.

In the PBN’s final discipline classification, the original 102 disciplines are merged into 29 joint evaluation groups (JEGs), 11 of which belong to SSH fields. Different scientific unit types within the same science fields are assigned to the same JEGs according to discipline. For example, all Polish history faculties and history institutes at the Polish Academy of Sciences are classified into one JEG called “History.”

As shown in Table 2, we compared the Flemish and the Polish SSH classifications and found four congruent disciplines. Of the four, a direct comparative analysis is possible for three disciplines: “Economics and Business,” “History,” and “Law.” Two Flemish categories, “Philosophy” and “Theology,” can be merged and compared with the Polish classification “Philosophy and Theology” to form the fourth discipline for comparison in this study.
Table 1. Publication types and inclusion criteria in the VABB–SHW and the PBN

<table>
<thead>
<tr>
<th>VABB–SHW</th>
<th>Inclusion criteria</th>
<th>PBN</th>
<th>Inclusion criteria</th>
</tr>
</thead>
</table>
| Article           |                    | Article | 1. Articles in journals indexed on the Polish Journal Ranking prepared by the Ministry of Science and Higher Education. This ranking organizes journals into three lists—A, B, and C:  
|                   |                    |       | • The *A list*: Journals indexed in the Journal Citation Reports;  
|                   |                    |       | • The *B list*: Polish (and until 2014, also foreign) journals without an impact factor;  
|                   |                    |       | • The *C list*: Journals indexed in the European Reference Index for the Humanities.  
|                   |                    |       | 2. Articles in foreign journals written in a foreign language (at least half an author sheet length). |
| Monograph         |                    | Monograph |  |
| Edited book       |                    | Edited book |  |
| Conference proceeding |                |  |
| Book chapter      |                    |  |
Table 2. Matching the VABB–SHW and PBN SSH discipline classifications

<table>
<thead>
<tr>
<th>VABB–SHW disciplines</th>
<th>PBN JEGs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archeology</td>
<td>Arts</td>
</tr>
<tr>
<td>Art history</td>
<td>Language, bibliography and culture studies</td>
</tr>
<tr>
<td>Communications studies</td>
<td>Music</td>
</tr>
<tr>
<td>History</td>
<td>History</td>
</tr>
<tr>
<td>Law</td>
<td>Law</td>
</tr>
<tr>
<td>Linguistics</td>
<td>Performing arts</td>
</tr>
<tr>
<td>Literature</td>
<td></td>
</tr>
<tr>
<td>Philosophy</td>
<td></td>
</tr>
<tr>
<td>Theology</td>
<td></td>
</tr>
<tr>
<td>Criminology</td>
<td>Plastic arts</td>
</tr>
<tr>
<td>Economics and business</td>
<td>Economics and business</td>
</tr>
<tr>
<td>Educational sciences</td>
<td>Social sciences</td>
</tr>
<tr>
<td>Political science</td>
<td>Theater</td>
</tr>
<tr>
<td>Psychology</td>
<td></td>
</tr>
<tr>
<td>Social health sciences</td>
<td></td>
</tr>
<tr>
<td>Sociology</td>
<td></td>
</tr>
</tbody>
</table>

Materials and Methods

For the purposes of this paper, we analyzed data for the SSH literature as a whole, as well as for specific SSH disciplines. For our analysis of SSH publication patterns in Flanders and Poland, we used VABB–SHW and PBN data collected between 2009 and 2014. The two datasets we used are described below:

(A) Dataset A contained bibliographic information for 77,870 publications registered in the VABB–SHW and 134,111 publications registered in the PBN between 2009 and 2014. For our analysis of overall SSH publication patterns in Flanders and Poland, we used the number of publications per country as our unit of analysis, along with two nominal sub-variables, i.e., (1) publication type (article, monograph, edited book, or chapter) and (2) language (local language [Dutch in Flanders, Polish in Poland], English, or other), as well as one rank variable, i.e., the publication year (2009–2014). Dataset A included no duplicate publications at the country level.

(B) Dataset B contained bibliographic information for publications within four disciplines in Flanders and Poland, including 37,338 “Economics and Business” publications; 21,345 “History” publications; 32,285 “Law” publications; and 12,333 “Philosophy and Theology” publications. For our discipline-level analysis of publication patterns in Flanders and Poland, we used the number of publications per discipline as our unit of analysis, along with the same set of variables used in dataset A. In the VABB–SHW, publications can be assigned to more than one discipline, which means that publications assigned to disciplines in the Flemish database do not constitute distinct sets. However, in dataset B, the percentage of publications with only one assigned discipline was high, including for “Economics and Business” (92.28%), “History” (95.85%), “Law” (94.12%), and “Philosophy and Theology” (96.86%). Publications assigned to two or more disciplines were excluded from our analysis ($N = 1,042$).

Results

In the first part of this section, we describe the language and publication type occurrence frequencies in Flanders and Poland based on dataset A. We analyze the variability in the
proportion of publications in English and in article form between 2009 and 2014 in Flanders and Poland separately. In the second part of this section, we present the differences in the use of publication types and languages between SSH disciplines in Flanders and Poland based on dataset B.

Part A: Characteristics of SSH Publications Patterns in Flanders and Poland

Table 3 shows the distribution of publication types in Flanders and Poland. Though there are country-level differences for all publication types, the highest disproportion is between articles and chapters. In Flanders, articles are the most common publication type (57%), whereas in Poland, articles constitute only 26.2% of all published work. At the same time, in Flanders, chapters constitute only 32.6% of all published work, whereas in Poland, chapters constitute 55.9%.

Table 3. Number and percentage of SSH publications per type between 2009 and 2014

<table>
<thead>
<tr>
<th></th>
<th>Flanders (N)</th>
<th>Flanders (%)</th>
<th>Poland (N)</th>
<th>Poland (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article</td>
<td>44,419</td>
<td>57.0</td>
<td>35,091</td>
<td>26.2</td>
</tr>
<tr>
<td>Monograph</td>
<td>4,445</td>
<td>5.7</td>
<td>14,611</td>
<td>10.9</td>
</tr>
<tr>
<td>Edited book</td>
<td>3,652</td>
<td>4.7</td>
<td>9,565</td>
<td>7.1</td>
</tr>
<tr>
<td>Chapter</td>
<td>25,354</td>
<td>32.6</td>
<td>74,844</td>
<td>55.8</td>
</tr>
<tr>
<td>Total</td>
<td>77,870</td>
<td>100.0</td>
<td>134,111</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Figure 1 displays the distribution of publication types between 2009 and 2014. In Flanders, we observed that patterns related to publication type were rather stable, whereas in Poland, we observed considerable changes of the proportions of publication types. In 2013, the most significant change occurred in Poland: the share of the articles dramatically increased. Whereas in 2009, articles had constituted 19.3% of all publications in Poland, by 2013 this proportion had increased to 44.6%. In Flanders, in contrast, articles constituted 55.4% of the total in 2009 and 57.6% in 2013.

---

Figure 1. Proportion of publication types in Flanders and Poland
Table 4 presents the most commonly used publication languages in Flanders and Poland. On the one hand, in Flanders, most publications were written in English (55.7%), and 35.8% were written Dutch, the local language. On the other hand, in Poland, only a small percentage of work was written in English (11.8%), and the local language, Polish, dominated academic work (82.7%). Both in Flanders and in Poland, a limited share of publications appeared in languages other than English or the local language (8.5% and 5.5%, respectively).

Table 4. Number and percentage of SSH publications per language between 2009 and 2014

<table>
<thead>
<tr>
<th>Language</th>
<th>Flanders (N)</th>
<th>Flanders (%)</th>
<th>Poland (N)</th>
<th>Poland (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local</td>
<td>27,863</td>
<td>35.8</td>
<td>110,871</td>
<td>82.7</td>
</tr>
<tr>
<td>English</td>
<td>43,364</td>
<td>55.7</td>
<td>15,866</td>
<td>11.8</td>
</tr>
<tr>
<td>Other</td>
<td>6,643</td>
<td>8.5</td>
<td>7,374</td>
<td>5.5</td>
</tr>
<tr>
<td>Total</td>
<td>77,870</td>
<td>100.0</td>
<td>134,111</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Figure 2 shows the distribution of publication languages between 2009 and 2014. The proportion of the publications in English is much higher in Flanders than in Poland. However, in both Flanders and Poland, there was an increase in the proportion of English publications. The difference in the proportion of English publications in 2009 versus 2013 was significant in both Flanders and Poland.

Figure 2. Publications in English, the local language, and other languages as a percentage of the total (all publication types)

Part B: Characteristics of the Publication Patterns According to Four Disciplines in Flanders and Poland

Table 5 shows the distribution of publication types within four disciplines: “Economics and Business,” “History,” “Law,” and “Philosophy and Theology.” The proportion of articles differed according to disciplines: articles were the dominant publication type within “Economics and Business,” “Law,” and “Philosophy and Theology” in Flanders, while within “History,” articles and chapters were both common. In Poland, chapters were the dominant publication type within each of the four disciplines.
Table 5. SSH publications per type and by discipline in Flanders and Poland between 2009 and 2014

<table>
<thead>
<tr>
<th>Publication type</th>
<th>Economics and Business</th>
<th>History</th>
<th>Law</th>
<th>Philosophy and Theology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FLA</td>
<td>POL</td>
<td>FLA</td>
<td>POL</td>
</tr>
<tr>
<td>N</td>
<td>5,833</td>
<td>9,264</td>
<td>1,681</td>
<td>3,945</td>
</tr>
<tr>
<td>%</td>
<td>63.5</td>
<td>32.9</td>
<td>42.6</td>
<td>22.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Publication type</th>
<th>Economics and Business</th>
<th>History</th>
<th>Law</th>
<th>Philosophy and Theology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FLA</td>
<td>POL</td>
<td>FLA</td>
<td>POL</td>
</tr>
<tr>
<td>N</td>
<td>563</td>
<td>2,474</td>
<td>293</td>
<td>2,127</td>
</tr>
<tr>
<td>%</td>
<td>6.1</td>
<td>8.8</td>
<td>7.4</td>
<td>12.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Publication type</th>
<th>Economics and Business</th>
<th>History</th>
<th>Law</th>
<th>Philosophy and Theology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FLA</td>
<td>POL</td>
<td>FLA</td>
<td>POL</td>
</tr>
<tr>
<td>N</td>
<td>269</td>
<td>2,102</td>
<td>230</td>
<td>1,172</td>
</tr>
<tr>
<td>%</td>
<td>2.9</td>
<td>7.5</td>
<td>5.8</td>
<td>6.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Publication type</th>
<th>Economics and Business</th>
<th>History</th>
<th>Law</th>
<th>Philosophy and Theology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FLA</td>
<td>POL</td>
<td>FLA</td>
<td>POL</td>
</tr>
<tr>
<td>N</td>
<td>2,521</td>
<td>14,312</td>
<td>1,746</td>
<td>10,151</td>
</tr>
<tr>
<td>%</td>
<td>27.4</td>
<td>50.8</td>
<td>44.2</td>
<td>58.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Publication type</th>
<th>Economics and Business</th>
<th>History</th>
<th>Law</th>
<th>Philosophy and Theology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FLA</td>
<td>POL</td>
<td>FLA</td>
<td>POL</td>
</tr>
<tr>
<td>N</td>
<td>9,186</td>
<td>28,152</td>
<td>3,950</td>
<td>17,395</td>
</tr>
<tr>
<td>%</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Note: FLA (Flanders); POL (Poland)

Table 6 presents the proportion of publications in English, the local language, and other languages in Flanders and Poland, according to the four defined disciplines.

Table 6. Languages per publication type in the SSH literature by discipline in Flanders and Poland between 2009 and 2014

<table>
<thead>
<tr>
<th>Publication type</th>
<th>Economics and Business</th>
<th>History</th>
<th>Law</th>
<th>Philosophy and Theology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FLA</td>
<td>POL</td>
<td>FLA</td>
<td>POL</td>
</tr>
<tr>
<td>N</td>
<td>1,807</td>
<td>22,899</td>
<td>1,699</td>
<td>14,047</td>
</tr>
<tr>
<td>%</td>
<td>19.7</td>
<td>81.3</td>
<td>43.0</td>
<td>80.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Publication type</th>
<th>Economics and Business</th>
<th>History</th>
<th>Law</th>
<th>Philosophy and Theology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FLA</td>
<td>POL</td>
<td>FLA</td>
<td>POL</td>
</tr>
<tr>
<td>N</td>
<td>7,114</td>
<td>5,065</td>
<td>1,658</td>
<td>2,271</td>
</tr>
<tr>
<td>%</td>
<td>77.4</td>
<td>18.0</td>
<td>42.0</td>
<td>13.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Publication type</th>
<th>Economics and Business</th>
<th>History</th>
<th>Law</th>
<th>Philosophy and Theology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FLA</td>
<td>POL</td>
<td>FLA</td>
<td>POL</td>
</tr>
<tr>
<td>N</td>
<td>265</td>
<td>188</td>
<td>593</td>
<td>1,077</td>
</tr>
<tr>
<td>%</td>
<td>2.9</td>
<td>0.7</td>
<td>15.0</td>
<td>6.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Publication type</th>
<th>Economics and Business</th>
<th>History</th>
<th>Law</th>
<th>Philosophy and Theology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FLA</td>
<td>POL</td>
<td>FLA</td>
<td>POL</td>
</tr>
<tr>
<td>N</td>
<td>9,186</td>
<td>28,152</td>
<td>3,950</td>
<td>17,395</td>
</tr>
<tr>
<td>%</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Note: FLA (Flanders); POL (Poland)

The proportion of publications in English varied by discipline. Within “Economics and Business” and “Philosophy and Theology,” English was the dominant publication language in Flanders. Within “History,” the proportions of publications in English and Dutch were similar. However, Dutch was the dominant publication language within “Law” in Flanders. Polish, the local language in Poland, was dominant within all four disciplines in Poland.

Conclusion

As we saw in the case of Norway (Sivertsen 2014), in Flanders, it appears that SSH publication patterns related to publication language and type evolved gradually between 2009
and 2014. This observation contrasts with what we observed in Poland, where the proportions of publication types changed significantly between 2009 and 2014 and the use of English only gradually increased. We furthermore observed that the variability over time in the proportions of publications in English and in article form were similar for both the SSH literature as a whole and each of four disciplines separately.

These publication patterns are rooted in scholarly traditions (Whitley, 2000; Ziman, 1968), as the relatively stable cases of Flanders and Norway show. Nonetheless, our analysis has revealed that discipline-level publication patterns differ more across countries than SIvertsen (2016) and van Leeuwen (2006) initially suggested. It appears that similarities between disciplines depend on not only the analogies within disciplines but also the similarities between countries. In comparing Flemish and Polish SSH publication patterns, we observed few similarities that could be compared at either the aggregate level or the discipline level. In each of the four Flemish disciplines, articles were more dominant than in the Polish disciplines, and the share of English publications was significantly higher among Flemish scholars compared to Polish scholars.

In general, our findings revealed two publication pattern characteristics that must be interpreted in a broader context. The first is related to the proportions of publication types, which were relatively stable in Flanders but were subject to significant change in Poland. Since 2009, in Poland, the number of journal articles has increased, whereas the number of book chapters has decreased. The number of scholarly books has decreased significantly as well. An interpretation of these changes is possible when we identify the main underlying mechanisms. In Poland, the regulations for both the performance-based research funding system and for academic promotions changed considerably between 2009 and 2014 (Kulczycki, 2017). Furthermore, science policy in Poland has increasingly provided incentives for publishing articles and for publishing in English. As our findings show, these policies seem have achieved some of their intended effects. However, we have not analyzed the quality of the articles, for example, in terms of publication channel (e.g., top-tier journals indexed in WoS or Scopus). In other words, whether unintended effects have occurred as well remains to be studied.

The other characteristic is related to publication language, and particularly publications written in English. Belgium (Flanders) and Poland are non-English speaking countries. However, in terms of the scholars working there, internationalization in Flanders and Poland is different from other countries. An even more important explanatory factor, however, might be their respective XX-century histories. In Poland, Russian was a compulsory language at school prior to 1989, and publishing in English was not the best way to communicate research results. Moreover, presently, many Polish scholars perceive academic publishing in English as a form of communicative inequality that promotes “linguistic injustice” (Hyland 2016).

Acknowledgments
The authors would like to thank Ewa A. Rozkosz for her useful suggestions and remarks. The work of EK was supported by the National Programme for the Development of Humanities in Poland [grant number 0057/NPHR3/H11/82/2014]. The authors are indebted to COST Action CA1537 “European Network for Research Evaluation in the Social Sciences and the Humanities” for supporting this work.

References


Assessment Criteria for Early Career Researcher’s Proposals in the Humanities

Michael Ochsner¹  Sven E. Hug²  Hans-Dieter Daniel³

¹ochsner@gess.ethz.ch
ETH Zurich, Zurich (Switzerland)
FORS, Lausanne (Switzerland)

²sven.hug@gess.ethz.ch
ETH Zurich, Zurich (Switzerland)
University of Zurich, Zurich (Switzerland)

³daniel@gess.ethz.ch
ETH Zurich, Zurich (Switzerland)
University of Zurich, Zurich (Switzerland)

Abstract
Competitive research grants become more and more important in the careers of young scholars. If grants are making careers, the decision for the grant winners is important and needs to be fair, consistent and transparent. In this research in progress paper, we present evaluation criteria for research proposals from early career researchers in the humanities. We apply a bottom-up procedure to identify evaluation criteria that reach consensus among the humanities scholars themselves. We identified 23 aspects pertaining to 9 criteria for the assessment of research proposals. There are no differences between the selection of aspects that reach consensus among the scholars regarding whether the applicant is a doctoral student or a postdoc, nor did we find differences in the selection of aspects between disciplines. We found slight differences in the ratings between tenured and non-tenured scholars and between women and men. Tenured scholars and women each emphasized an additional aspect.

Conference Topic
Science Policy and Research Evaluation

Introduction
Competitive research grants become more and more important in the careers of young scholars (van Arensbergen, van der Weijden, & van den Besselaar, 2014a, b). The acquisition of such grants is seen as a sign of quality of scholarship by senior researchers (see e.g. van Arensbergen et al., 2014a, b) as well as in evaluation procedures (see e.g. Ochsner et al., 2012). This process is closely linked to the shift to the notion of excellence in higher education policy. If a higher education institution adheres to the notion of excellence, it has to recruit excellent scholars. Therefore, governments and universities focus on talent selection processes and increase the support for early career researchers. Research policy implements ‘excellence’ amongst others through competitive research funding and temporary positions for early and mid-career researchers (van Arensbergen et al., 2014b; van den Akker, 2016). While competitive funding first concerned Science, Technology, Engineering and Medicine (STEM) disciplines due their need for expensive infrastructure and large teams (Krull and Tepperwien, 2016), in the humanities the acquisition of competitive grants has not been very important in the past. However, the change to a focus on talent selection and temporary employment applies also to the humanities. Therefore, the assessment and selection of research proposals become more and more important in the humanities as well.

While there are studies on peer and panel review, they focus mainly on selection biases and fairness (see e.g. Bornmann, Mutz & Daniel, 2008; Bornmann, Mutz, Marx, Schier &
Daniel, 2011; Lamont, 2009). There is a lack of knowledge about what quality of research proposals means and how it can be identified, especially so in the humanities (see e.g. Hemlin, 1993). Little is known about what criteria peers have in mind when evaluating a proposal and even less how they weight these criteria. However, judging a work without criteria is inconsistent and not adequate for judging merit, as Thorngate, Dawes and Foddy (2009) conclude their comprehensive research on decision making. They found that judging separately according to specified criteria reveals more consistent results (Thorngate et al., 2009, p. 26). Such intra-rater reliability (in distinction of the inter-rater reliability) is of utter importance when judging merit (i.e. a reviewer would give the same rating for application A when rating application A before application B or after it; or give the same rating at a later point in time). It is not only crucial to have reliable judgments as a basis in the review process. When deciding for future careers, it is also important that the criteria are explicit and clear so that young scholars do know what to deliver and, in case of a negative evaluation, how to improve. Furthermore, explicit criteria serve transparency. All these points are important for the judgment of merit to be fair and consistent (Thorngate et al., 2009). The growing importance of research grants for the further career of young scholars makes it particularly important that the best applicants are awarded the grant. Therefore, an adequate procedure for selecting the best proposals must be applied.

In this research in progress paper, we present quality criteria for the ex-ante assessment of research proposals from early career researchers in the humanities. Applying a bottom-up approach we base the evaluation criteria on scholars’ ratings of quality criteria regarding their adequacy for the use in such an assessment situation. Particularly, we are investigating the following research questions: a) are there differences between the criteria for evaluating the proposals from PhD students and those for evaluating proposals from postdocs? b) is there a common set of quality criteria across disciplines that can be used to adequately judge research proposals? c) do tenured professors emphasize other criteria than the young scholars themselves? d) are there gender differences regarding the preferences for criteria?

This paper is organized in the following way: First, we present the approach and methods used to identify the important quality criteria for a specific evaluation situation. We then present the methods with which we investigate differences in the preferences for quality criteria between sub-groups of our sample: level of the grant (PhD or postdoc), discipline, academic status and gender. We finally conclude regarding the differences between sub-groups of the sample reflecting the generalizability of our results.

Methods and data used

For the selection of the criteria to be included in the evaluation sheet, we designed a questionnaire containing nine criteria, specified by 23 aspects, for judging the quality and potential of the research proposal. We draw from our previous research in which we developed a catalogue of criteria for research quality in the humanities in a strictly bottom-up procedure, i.e. the scholars formulated their own criteria during several steps and using different methods (Ochsner, Hug & Daniel, 2014). Based on this catalogue, we selected, adapted and expanded the criteria to the evaluation situation of ex-ante research proposal assessment of early career researchers and added some criteria usually used in such evaluation situations (i.e., information about the applicant). The questionnaire was sent out to all Swiss scholars holding a doctoral degree in the humanities (theology/religious studies were excluded due to another project fielding a similar survey at the same time in these disciplines). The scholars were to rate the criteria for their suitability for the evaluation of research proposals. To do so, they had to give their agreement on a 6-point scale with a statement that consisted of a generic part and a part specific to the aspect that is rated. The generic part read: ‘A project application is assessed appropriately, if the assessment considers whether...’ while the

Using the ratings, we identified criteria and aspects that reach consensus among the scholars. An aspect reaches consensus when it is clearly approved by the majority (at least 50% of the scholars rate the aspect with at least a ‘5’) and only a small majority disapproves the aspect (less than 10% rate the aspect negatively, that is with a ‘1’, ‘2’ or ‘3’).

To answer our research questions, we identified differences in the ratings of the aspects between different sub-groups of our sample using standardized effect sizes, i.e. Cohen’s d (Cohen, 1988). We applied a threshold of Cohen’s $d=0.2$ as suggested by Cohen (1988), who proposes the rule of thumb that $d>0.2$ equals a small effect, $d>0.5$ a medium effect, and $d>0.8$ a large effect. Because our sample is a full population survey (i.e., all humanities scholars in Switzerland were invited to participate) and not a random sample, we cannot use inferential statistics but use bootstrap resampling with 1,000 replications to estimate the stability of the results instead, using bootstrapped 95% stability intervals. This serves to account for the effect of possible outliers and as a measure of stability because not all humanities scholars in the population respond to the survey (see Schneider & van Leeuwen, 2014).

**Results**

The questionnaire was sent out to 2,609 humanities scholars of whom 916 filled in the questionnaire. This amounts to an overall response rate of 35%, which is a very good response rate compared to similar survey projects (see e.g. Braun & Ganser, 2011; Cardoso, Rosa & Santos, 2013; Giménez-Toledo & Román-Román, 2013). The response rate differed between disciplines in the sense that scholars from law studies participated to a quiet smaller degree: While the response rate in language and linguistics amounted to 38% and in history and cultural studies to 41%, the response rate among law scholars was at 24%. This can be explained by the fact that in law studies, our sample contained a significant number of persons that are primarily active as lawyers or judges and teach irregularly at the university. Those scholars did not participate (many of them sent us an email with the excuse that they felt it was not appropriate for them to answer the questionnaire due to a certain distance from academic life).

From the 23 aspects assigned to 9 criteria that the scholars rated, 13 aspects pertaining to 6 criteria reached consensus (see table 1). All aspects that reached the threshold of not more than 10% negative ratings also reached a median of 5. Therefore, we only list the results for the percentage of negative ratings as this was the decisive criterion. As the bootstrapped stability intervals show, the results are quite stable.

The important criteria for the evaluation of research proposals of young humanities scholars are their originality, feasibility, rigour, relevance, complexity and variety. Originality is defined by the aspects identifying gaps in existing knowledge, using innovative data, presenting new findings. Feasibility is defined by a realistic timetable and resources. Rigour is defined by the aspects appropriate research process, expression of the state of research and choice of method as well as a stringent argumentation and understandable language. Relevance is defined as academic relevance. Complexity is defined as making complexity visible. Variety is defined as contribution to the variety of research.

Looking at the Cohen’s d, we find little differences in general. Almost all coefficients are below the threshold of $d=0.2$. Between the assessment of proposals from doctoral students and postdocs, only two aspects reach a $d>0.2$: independence ($d=0.20$) and the applicant’s publication list ($d=0.30$) are rated less favourably for the assessment of proposals from doctoral students than for those from postdocs. However, for both groups, the two aspects do
not reach consensus. Therefore, the selection of criteria for the assessment of proposals from doctoral students is the same as for those from postdocs.

Table 1. Overall mean, percentage of negative ratings (with bootstrapped 95% stability intervals in parentheses), and Cohen’s d of subgroups for the ratings of the aspects

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Mean</th>
<th>% of neg. ratings</th>
<th>Cohen’s d Doc vs Postdoc</th>
<th>Cohen’s d Lang. vs HistCult</th>
<th>Cohen’s d Law. vs. Lang.</th>
<th>Cohen’s d HistCult vs. Law</th>
<th>Cohen’s d Tenure</th>
<th>Cohen’s d Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independence</td>
<td>4.77</td>
<td>0.16</td>
<td>-0.20</td>
<td>-0.20</td>
<td>0.08</td>
<td>0.12</td>
<td>-0.08</td>
<td>0.10</td>
</tr>
<tr>
<td>(4.69-4.85)</td>
<td></td>
<td>(0.14-0.19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Originality: Identify Gaps</td>
<td>5.31</td>
<td>0.03</td>
<td>-0.07</td>
<td>0.10</td>
<td>-0.18</td>
<td>0.07</td>
<td>-0.07</td>
<td>0.15</td>
</tr>
<tr>
<td>(5.25-5.36)</td>
<td></td>
<td>(0.02-0.04)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Originality: Innovative Data</td>
<td>5.18</td>
<td>0.04</td>
<td>-0.01</td>
<td>-0.10</td>
<td>-0.12</td>
<td>0.23</td>
<td>0.11</td>
<td>0.13</td>
</tr>
<tr>
<td>(5.12-5.23)</td>
<td></td>
<td>(0.03-0.05)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Originality: New Research Topic</td>
<td>4.75</td>
<td>0.12</td>
<td>0.02</td>
<td>0.00</td>
<td>0.10</td>
<td>-0.10</td>
<td>0.07</td>
<td>-0.01</td>
</tr>
<tr>
<td>(4.68-4.83)</td>
<td></td>
<td>(0.10-0.14)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Originality: New Approach</td>
<td>4.80</td>
<td>0.13</td>
<td>0.02</td>
<td>0.04</td>
<td>-0.29</td>
<td>0.25</td>
<td>-0.03</td>
<td>0.13</td>
</tr>
<tr>
<td>(4.72-4.87)</td>
<td></td>
<td>(0.10-0.15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Originality: New Paradigm</td>
<td>4.53</td>
<td>0.19</td>
<td>0.01</td>
<td>0.05</td>
<td>0.08</td>
<td>-0.13</td>
<td>-0.05</td>
<td>-0.02</td>
</tr>
<tr>
<td>(4.45-4.62)</td>
<td></td>
<td>(0.16-0.21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Originality: New Finding</td>
<td>4.99</td>
<td>0.08</td>
<td>0.05</td>
<td>0.07</td>
<td>0.06</td>
<td>-0.13</td>
<td>-0.19</td>
<td>0.16</td>
</tr>
<tr>
<td>(4.93-5.05)</td>
<td></td>
<td>(0.06-0.09)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feasibility: Timetable</td>
<td>5.03</td>
<td>0.07</td>
<td>-0.03</td>
<td>-0.05</td>
<td>-0.04</td>
<td>0.08</td>
<td>0.19</td>
<td>0.09</td>
</tr>
<tr>
<td>(4.96-5.09)</td>
<td></td>
<td>(0.05-0.08)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feasibility: Resources</td>
<td>5.10</td>
<td>0.05</td>
<td>-0.01</td>
<td>-0.05</td>
<td>0.02</td>
<td>0.03</td>
<td>0.16</td>
<td>0.02</td>
</tr>
<tr>
<td>(5.04-5.16)</td>
<td></td>
<td>(0.03-0.06)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rigour: Research Process</td>
<td>5.31</td>
<td>0.03</td>
<td>-0.06</td>
<td>-0.07</td>
<td>0.04</td>
<td>0.02</td>
<td>0.29</td>
<td>0.09</td>
</tr>
<tr>
<td>(5.26-5.36)</td>
<td></td>
<td>(0.02-0.04)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rigour: State of Research</td>
<td>5.29</td>
<td>0.03</td>
<td>0.01</td>
<td>-0.05</td>
<td>-0.17</td>
<td>0.23</td>
<td>0.20</td>
<td>0.13</td>
</tr>
<tr>
<td>(5.23-5.35)</td>
<td></td>
<td>(0.02-0.05)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rigour: Choice of Method</td>
<td>4.98</td>
<td>0.06</td>
<td>0.03</td>
<td>-0.09</td>
<td>0.05</td>
<td>0.04</td>
<td>0.20</td>
<td>0.15</td>
</tr>
<tr>
<td>(4.92-5.04)</td>
<td></td>
<td>(0.04-0.07)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rigour: Argumentation</td>
<td>5.54</td>
<td>0.02</td>
<td>-0.03</td>
<td>-0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.23</td>
<td>0.16</td>
</tr>
<tr>
<td>(5.49-5.59)</td>
<td></td>
<td>(0.01-0.03)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rigour: Understandable</td>
<td>5.51</td>
<td>0.02</td>
<td>-0.01</td>
<td>-0.15</td>
<td>0.18</td>
<td>-0.04</td>
<td>0.16</td>
<td>-0.08</td>
</tr>
<tr>
<td>(5.46-5.55)</td>
<td></td>
<td>(0.01-0.02)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevance: Academia</td>
<td>5.06</td>
<td>0.07</td>
<td>0.01</td>
<td>-0.11</td>
<td>0.05</td>
<td>0.06</td>
<td>0.19</td>
<td>-0.08</td>
</tr>
<tr>
<td>(4.99-5.13)</td>
<td></td>
<td>(0.06-0.09)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevance: Societal</td>
<td>3.81</td>
<td>0.36</td>
<td>-0.02</td>
<td>-0.22</td>
<td>0.72</td>
<td>-0.50</td>
<td>-0.13</td>
<td>0.08</td>
</tr>
<tr>
<td>(3.72-3.90)</td>
<td></td>
<td>(0.33-0.39)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cultural Heritage</td>
<td>4.46</td>
<td>0.19</td>
<td>-0.03</td>
<td>-0.06</td>
<td>-0.60</td>
<td>0.65</td>
<td>-0.12</td>
<td>0.02</td>
</tr>
<tr>
<td>(4.38-4.55)</td>
<td></td>
<td>(0.17-0.22)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complexity</td>
<td>4.98</td>
<td>0.08</td>
<td>-0.15</td>
<td>-0.12</td>
<td>-0.06</td>
<td>0.18</td>
<td>0.01</td>
<td>0.14</td>
</tr>
<tr>
<td>(4.92-5.05)</td>
<td></td>
<td>(0.06-0.10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variety</td>
<td>4.96</td>
<td>0.08</td>
<td>-0.11</td>
<td>-0.05</td>
<td>-0.05</td>
<td>0.10</td>
<td>-0.11</td>
<td>0.20</td>
</tr>
<tr>
<td>(4.89-5.02)</td>
<td></td>
<td>(0.06-0.10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person: CV</td>
<td>4.79</td>
<td>0.11</td>
<td>-0.16</td>
<td>-0.10</td>
<td>0.11</td>
<td>-0.02</td>
<td>0.34</td>
<td>0.15</td>
</tr>
<tr>
<td>(4.71-4.86)</td>
<td></td>
<td>(0.09-0.13)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person: Diploma</td>
<td>4.58</td>
<td>0.12</td>
<td>-0.12</td>
<td>0.02</td>
<td>0.14</td>
<td>-0.17</td>
<td>0.23</td>
<td>0.21</td>
</tr>
<tr>
<td>(4.51-4.65)</td>
<td></td>
<td>(0.10-0.14)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person: Publications</td>
<td>4.52</td>
<td>0.16</td>
<td>-0.30</td>
<td>-0.01</td>
<td>0.02</td>
<td>-0.01</td>
<td>0.39</td>
<td>0.02</td>
</tr>
<tr>
<td>(4.45-4.60)</td>
<td></td>
<td>(0.14-0.19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person: Recommendations</td>
<td>3.90</td>
<td>0.29</td>
<td>0.02</td>
<td>0.07</td>
<td>-0.06</td>
<td>-0.01</td>
<td>0.10</td>
<td>0.19</td>
</tr>
<tr>
<td>(3.81-3.98)</td>
<td></td>
<td>(0.26-0.32)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Doc=proposals from doctoral students; postdoc=proposals from postdocs; Lang.=language and literature; HistCult=history and cultural studies; Law=law studies
Regarding disciplines, we find six aspects reaching a Cohen’s $d>0.2$ (i.e. independence, innovative data, new approach, state of research, cultural heritage and societal relevance), but only two aspects, for which ratings are different to a greater degree: the criterion cultural heritage is rated much higher in the disciplines language and literature as well as history and cultural studies than in law studies ($d=0.6$ and $d=0.65$), while societal relevance is rated much higher in law studies ($d=0.7$ and $d=0.5$). But again, the aspects do not differ regarding consensus. Thus, while we find disciplinary differences regarding the ratings of some aspects, the selection of aspects that reach consensus does not change between disciplines.

We also find differences between the ratings from tenured and non-tenured scholars. From the criterion rigour, tenured scholars rate all aspects bar one higher than non-tenured scholars: research process ($d=0.29$), state of research ($d=0.20$), choice of method ($d=0.20$) and argumentation ($d=0.23$). Furthermore, tenured scholars rate the applicant’s CV ($d=0.34$), his or her diploma ($d=0.23$) and his or her publication list ($d=0.39$) higher than non-tenured scholars. The applicant’s CV reaches consensus among tenured scholars as an important criterion for the assessment of young scholars’ research proposals while it does not reach consensus among non-tenured scholars. Other than that, the selection of aspects reaching consensus does not differ between tenured and non-tenured scholars.

Regarding gender, we only find two aspects that reach the threshold of Cohen’s $d=0.2$. Women rate the variety ($d=0.20$) and the applicant’s diploma ($d=0.21$) higher than men. Variety reaches consensus among women but not among men (the bootstrapped stability interval for men amounts to 0.07-0.12), revealing slight gender differences regarding the ratings of the criteria and aspects for the assessment of young scholars’ research proposals.

Conclusions

In this research in progress paper, we investigated the criteria and aspects that humanities scholars feel important and adequate to assess research proposals by young humanities scholars. We found that 23 aspects pertaining to 9 criteria reach consensus among the scholars: originality (identifying gaps, innovative data, new findings), feasibility (timetable and resources), rigour (research process, state of research, choice of method, stringent argumentation, understandable), relevance (academic relevance), complexity (making complexity visible) and variety (variety of research).

Regarding our first research question, whether there are differences between the criteria for evaluating the proposals from PhD students and those for evaluating proposals from postdocs, we found that there are no such differences. The same selection of criteria reached consensus for both evaluation situations. Regarding our second research question, whether there are disciplinary differences between the evaluation criteria, we also found that the same selection of criteria reached consensus in the three groups of disciplines we investigated. However, there were some differences in the means between disciplines that are related to the topics of research in the disciplines: scholars in language and literature and in history and cultural studies emphasised more the criterion cultural heritage than scholars in law studies, while scholars in law studies rated societal relevance higher. This points to the fact that while the same evaluation sheet can be used in all disciplines, the weighting of the criteria (can) differ by discipline. The third research question focused on differences between tenured and non-tenured scholars. We found very little differences, however, the applicant’s CV reached consensus among tenured scholars but not among non-tenured scholars. Finally, we investigated gender differences in the ratings. Regarding gender, the differences are rather small as well and only one criterion, variety of research, reached consensus among women but not among men. Thus, we can conclude that the 23 aspects pertaining to 9 criteria are rather generally applicable in the assessment of research proposals from early career researchers in the humanities.
Regarding the application of the criteria in the assessment of research proposals, it remains to be noted that we present findings from the perspective of the scholars themselves, thus referring to the academic quality of the proposal. In funding decisions, also criteria put forward by the funder might be added. In the further analysis of our data, we will delve deeper into the relation of the criteria for the assessment of proposals by early career researchers and general quality criteria identified in our previous research (Hug et al., 2013) as well as the interrelations of gender, tenure and disciplines regarding the ratings of the evaluation criteria.

Acknowledgements
The authors would like to thank swissuniversities for their grant for the project “Application of Bottom-up Criteria in the Assessment of Grant Proposals of Junior Researchers” within the “Programme P-3 Performances de la recherche en sciences humaines et sociales”. Matching funds were provided by the University of Zurich.

References


Cognitive and organizational classification of publications in the Social Sciences and Humanities

Raf Guns¹  Tim C.E. Engels¹  Frederik T. Verleysen¹

¹raf.guns@uantwerpen.be, tim.engels@uantwerpen.be, frederik.verleysen@uantwerpen.be
Centre for R&D Monitoring (ECOOM), Faculty of Social Sciences, University of Antwerp, Antwerp (Belgium)

Abstract
We study the discrepancy between two ways of classifying publications in the social sciences and humanities (SSH): on the basis of the journal’s contents (cognitive) or on the basis of the organizational structure of departments, faculties etc. (organizational). Using data from 43,191 journal articles (co-) authored by Flemish SSH researchers in 2005–2014, we compare the organizational classification with a cognitive classification based on OECD Fields Of Science. In virtually all cases, the cognitive discipline with the most publications is the one that most closely matches the organizational discipline, although there are great differences between disciplines. At a higher aggregation level, we find that 70% of publications from the humanities are published in humanities journals, while this is only the case for 53% of publications from the social sciences. Social sciences are shown to be closely linked to medicine and health sciences. The spread of publications over cognitive disciplines, including non-SSH disciplines, can be partially explained by the assignment of some journals to multiple FOS fields as well as multidisciplinary collaboration.

Conference Topic
Journals, databases and electronic publications; Science policy

Introduction
Publications, publication channels, and authors can be assigned to disciplines on the basis of different criteria. The most straightforward criterion is presumably the contents of the publications, but in practice other criteria are frequently used as well. Daraio and Glänzel (2016) distinguish between four types of classifications:
- cognitive: based on content of publications or publication channels;
- administrative: based on responsibility from the point of view of a government or research policy;
- organizational: based on internal organizational structure;
- qualification-based: based on education and competencies.
In this paper, we focus on the tension between organizational and cognitive classifications. Whitley (2000) has pointed out the “degree of separation of control” between the praxis of research and “administrative structures of work organizations.” In the case of Flanders, Belgium, for instance, the organizational structure of universities – departments, faculties, etc. – is traditionally mainly dictated by educational needs rather than research. One might expect that in practice some or all of these criteria tend to coincide, that is, that the four criteria will generally lead to a publication or its authors being classified in the same discipline(s). At the same time, it is clear that this is not absolute: for instance, a researcher with formal training in economics may carry out computer science research while working in a physics department. Relatedly, it has been shown that researchers from the same organizational discipline may be quite different in terms of publication patterns (Verleysen & Weeren, 2016a, 2016b).
Daraio and Glänzel (2016) give the example of the Flemish Academic Bibliographic Database for Social Sciences and Humanities (VABB-SHW), a comprehensive database of social
sciences and humanities (SSH) output in Flanders. The number of publications in the VABB-SHW is a parameter in the Flemish BOF-key university funding system (Verleysen, Ghesquière, & Engels, 2014). Contrary to Web of Science – another parameter in the BOF-key –, which uses a cognitive classification, VABB-SHW classifies publications according to the authors’ departments or research groups and is hence based on organizational criteria. The mapping of organizational structure to disciplines is carried out by the respective university’s research administration together with ECOOM-Antwerp, which oversees the development and maintenance of the VABB-SHW.

Daraio and Glänzel remark that the organizational classification “is acceptable for the purpose of funding allocation and in line with the original task of VABB, above all, because the necessary affiliation information is available also in the WoS data. However, for fine grained evaluation exercises the combination of the two classification types remains problematic.” In addition, a cognitive classification has the advantage of allowing easier comparison with the situation in other countries (Ossenblok, Engels, & Sivertsen, 2012; Pölönen et al., 2017).

In general, not much is known on the extent of the discrepancy between cognitive and organizational classifications. We explore this tension using data from the VABB-SHW. Specifically, we focus on the following questions:

1. To what extent is there a mismatch between cognitive and organizational classifications?
2. How is this mismatch different across disciplines and organizational units?
3. How fuzzy are the boundaries between the SSH and other fields of research?

Data and methods

Data were obtained from the VABB-SHW (https://www.ecoom.be/en/vabb). All 43,191 journal articles published in the 10-year period 2005–2014 and classified as peer-reviewed in view of the Flemish funding system were taken into account. These articles were published in, in total, 8,683 journals. If a journal had separate ISSNs for print and online versions, these were considered as the same journal.

Each journal was assigned to one or more disciplines from the OECD Fields Of Science or FOS (OECD, 2007). This was done as much as possible by relying on the journal’s classification in Web of Science, Scopus, the Norwegian CRIStiN database, and the UDC or DDC classification in the ISSN.org database. For each of these, a concordance table was used to translate to the FOS classification. A limited number of journals did not have a subject indication in any of the aforementioned data sources and was classified manually and independently by two people.

The following refinements were made to the FOS classification:

- History and archaeology was split into History and Archaeology;
- Languages and literature was split into Languages and linguistics and Literature;
- Philosophy, ethics and religion was split into Philosophy and ethics and Religion.

The rationale behind the refinement was to obtain cognitive classes that correspond more directly to VABB disciplines. Apart from the groupings Humanities general and Social sciences general, only the VABB discipline Criminology does not have a direct counterpart in the refined FOS scheme, where criminology is classified under Law. From the perspective of FOS, only Social and economic geography (besides Other humanities and Other social sciences) does not have a clear counterpart in VABB.

For each pair of VABB discipline and FOS field, we determine the number of publications assigned to both. Because the absolute numbers of publications vary between VABB disciplines and FOS fields, we use the share of publications, expressed as a percentage. This can be done from the perspective of a VABB discipline – share of publications from that discipline assigned to a FOS field compared to all publications from that discipline – or from the perspective of a
For the purposes of the present paper, we will mostly consider things from the perspective of VABB disciplines, i.e. the organizational classification of the publications. Note that publications can be assigned to more than one VABB discipline (in case of collaboration; 44%) as well as more than one FOS field (12%). Indeed, it seems plausible that interdisciplinary collaboration is one of the driving forces enabling researchers to go beyond their own organizational discipline (van Rijnsoever & Hessels, 2011). To assess the strength of this effect, we separately consider the set of publications assigned to only one VABB discipline and one FOS field (N=25430 or 59%).

In addition to visual representations using heat maps, we will quantify the concentration of a VABB discipline in FOS fields using the Gini index (Gini, 1921). This index is directly related to the Lorenz curve (Rousseau, 2007) and varies between 0 (perfect evenness) and 1 (perfect unevenness, i.e. concentration).

**Results**

First, we compare the share of publications from the six top-level FOS fields in the aggregated organizational fields of **Social sciences** and **Humanities**, using all 43,191 articles. This is summarized in Figure 1. In both cases, we see that the majority of publications is published in the corresponding FOS field, although the share in humanities (70%) is much larger than in social sciences (53%). Moreover, publications authored by researchers from the humanities appear in social science journals in 14%, whereas this is only 3% for the reverse case. The top-level FOS field of **Medical and health sciences** is the second largest recipient of publications from scholars affiliated to Social Sciences (30%).

![Figure 1](image-url). Heatmap of publications classified according to organizational unit (rows) and journal content (columns). Data are reported for top-level classifications and standardized per row.

Figure 2 provides a more detailed view. The order of disciplines is taken from the FOS classification (OECD, 2007); VABB disciplines mirror the order of the closest corresponding FOS fields. The rough ‘diagonal’ line in the right-hand side of the picture indicates that, overall, the organizational affiliation is at least an approximation of the intellectual content of publications. Gaps in the diagonal are explained by the fact that some FOS fields have no direct counterpart in VABB disciplines. Because the FOS field of **Health sciences** is classified under **Medical and health sciences**, the first row (Social health sciences) does not show up as part of the diagonal. We also point out that some pairs of disciplines, like linguistics and literature or philosophy and theology, are shown to be closely related.

The intensity of cells in the diagonal indicates the extent to which a VABB discipline corresponds to a FOS field. The disciplines with the highest correspondence are **Law** (79.7%), **Criminology** (74.0%, mapped to FOS field Law), and **Theology** (66.4%). On the other hand it can be seen that some disciplines have a much weaker disciplinary profile, such as **Sociology** (21.0%), **Social health sciences** (26.8%), and **Communication studies** (29.2%).

**Social health sciences** is the only organizational discipline that actually has a higher share of publications in another FOS field than the one we considered to be the closest corresponding: 32.1% of its publications appear in **Clinical medicine** journals. Several other VABB disciplines
also appear regularly in medical journals, like *Psychology*, *Sociology* (sociology of health), and *Philosophy* (ethical issues).

![Heatmap of publications classified according to organizational unit (rows) and journal content (columns). Data are standardized per row.](image1)

**Figure 2.** Heatmap of publications classified according to organizational unit (rows) and journal content (columns). Data are standardized per row.

![Difference in percentage points between shares based on single-assignment publications and shares based on all publications](image2)

**Figure 3.** Difference in percentage points between shares based on single-assignment publications and shares based on all publications

In the analysis presented above, a confounding factor may be the fact that some publications are assigned to multiple VABB disciplines and/or multiple FOS fields. In order to rule out this
factor, we also calculated the share of publications from FOS fields in each VABB discipline, taking only single-assignment publications into account, i.e. publications that are assigned to one VABB discipline and one FOS field. Comparing these shares with the ones obtained by taking all publications into account gives an indication of the importance of this factor and the ways it affects the results.

Figure 3 visualizes the difference between shares based on single-assignment publications and shares based on all publications. Red shades indicate that the percentage calculated using only single-assignment publications is higher than the one using all publications, whereas blue shades indicate that the percentage for single-assignment publications is lower. In general, the share based on single-assignment publications increases for the FOS field that corresponds best to a given VABB discipline and decreases for most of the other FOS fields. The exception is, again, Social health sciences, where we only see an increase for FOS field Clinical medicine. The strongest increase is observed for Archaeology (+21.3 p.p.), while the strongest decrease is found for Sociology’s involvement in FOS field Clinical medicine (−9.3 p.p.).

It is no coincidence that a strong effect can be observed for the field of archaeology, since this is typically a discipline that strongly relies on multidisciplinary collaboration with fields like chemistry and earth sciences as well as with other humanities disciplines. In other words, multidisciplinary collaboration should be considered an integral part of some, if not all, disciplines, and cannot be simply disregarded.

These findings clearly indicate that some organizational disciplines are relatively concentrated in terms of in which cognitive fields articles are classified, whereas others are more spread out. This can be quantified using a concentration measure like the Gini index (Gini, 1921). The results (Table 1) are related to, but not interchangeable with, the results regarding the match between FOS field and corresponding VABB discipline. While all disciplines are relatively concentrated over FOS fields, the difference between the most and the least concentrated ones – Law and History of arts respectively – is considerable. As one might expect, the Gini index that is based on single-assignment publications only is higher in all cases than the one based on all publications.

<table>
<thead>
<tr>
<th>VABB discipline</th>
<th>Gini index (all publications)</th>
<th>Gini index (single-assignment publications only)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Law</td>
<td>0.93</td>
<td>0.96</td>
</tr>
<tr>
<td>Criminology</td>
<td>0.92</td>
<td>0.96</td>
</tr>
<tr>
<td>Theology</td>
<td>0.92</td>
<td>0.94</td>
</tr>
<tr>
<td>Literature</td>
<td>0.90</td>
<td>0.92</td>
</tr>
<tr>
<td>Psychology</td>
<td>0.88</td>
<td>0.92</td>
</tr>
<tr>
<td>Social health sciences</td>
<td>0.88</td>
<td>0.92</td>
</tr>
<tr>
<td>Archaeology</td>
<td>0.87</td>
<td>0.94</td>
</tr>
<tr>
<td>Political sciences</td>
<td>0.86</td>
<td>0.91</td>
</tr>
<tr>
<td>History</td>
<td>0.84</td>
<td>0.91</td>
</tr>
<tr>
<td>Linguistics</td>
<td>0.82</td>
<td>0.91</td>
</tr>
<tr>
<td>Educational sciences</td>
<td>0.81</td>
<td>0.90</td>
</tr>
<tr>
<td>Sociology</td>
<td>0.77</td>
<td>0.84</td>
</tr>
<tr>
<td>Philosophy</td>
<td>0.77</td>
<td>0.84</td>
</tr>
<tr>
<td>Communication studies</td>
<td>0.76</td>
<td>0.84</td>
</tr>
<tr>
<td>Economics &amp; business</td>
<td>0.73</td>
<td>0.85</td>
</tr>
<tr>
<td>History of arts</td>
<td>0.71</td>
<td>0.87</td>
</tr>
</tbody>
</table>
Conclusions
Different types of classification do not always yield the same results. In this paper, we have examined the discrepancy between cognitive and organizational classifications in social sciences and humanities, using the example of the Flemish VABB-SHW database. Our results show that all organizational disciplines are to some extent prone to publishing ‘outside’ one’s own discipline but there are great differences between disciplines: while almost 80% of law publications end up in law journals, this is less than 30% for communication studies. At a higher aggregation level, we find that 70% of humanities publications are published in humanities journals, while this is only the case for 53% of social sciences publications. Social sciences are shown to be closely linked to medicine and health sciences; this is partially due to the inclusion of Social health sciences as a social science in the organizational classification but also to other social sciences like sociology. The spread of publications over cognitive disciplines, including non-SSH disciplines, can be partially explained by the assignment of some journals to multiple FOS fields as well as multidisciplinary collaboration.

Acknowledgments
This investigation has been made possible by the financial support of the Flemish government to ECOOM.

References


Peer review as a delineation criterion in data sources for the assessment and measurement of scholarly book publishing in social sciences and humanities

Elea Giménez-Toledo 1 Gunnar Sivertsen 2 Jorge Mañana-Rodríguez 3

1 elea.gimenez@cchs.csic.es
Spanish National Research Council (Spain)

2 gunnar.sivertsen@nifu.no
Nordic Institute for Studies in Innovation, Research and Education (NIFU)

3 jorge.mannana@cchs.csic.es
Spanish National Research Council (Spain)

Abstract:
This research in progress paper presents current research on the different approaches to the definition, documentation and application of peer review processes to scholarly books and book publishers. The documentation concerning how peer reviewed scholarly books are considered in the Norwegian, Finnish, Danish, Flemish and Spanish evaluation systems and national CRIS(s) has been analyzed and juxtaposed. The synthesis of the documentation includes the distinctive features of each definition of scholarly book’s peer review. Inclusion or exclusion of the evaluation system as well as the weight given to peer and non-peer-reviewed scholarly books are also analyzed. The existence of a common denominator to all systems seems plausible and would facilitate an eventual aggregation of different information systems. Such common denominator might be defined as the least number of common features across the different countries concerning how peer review in scholarly books is considered in their evaluation systems. Nevertheless, that common denominator might be more clearly identified once more evaluation systems are analyzed as per extracting general conclusions. This would entail coordination among researchers and transparency by publishers. It is concluded that peer review of scholarly books is a highly relevant feature of scholarly output; also, the need of basic research on scholarly books’ peer review practices becomes clear from the diversity of approaches observed. Further research would involve quantitative analysis of peer-reviewed publishers in the different countries.

Keywords: Scholarly book publishers, peer review, research evaluation, information systems / CRIS.

Conference Topic
Science communication

Introduction:
The Association of American University Presses (AAUP), which represents a global community of independent scholarly publishers, demands “high standards of editorial quality and peer review” from all its members. AAUP was recently able to complete a Handbook of Best Practices for Peer Review (AAUP, 2016). This is a sign that independent peer review is now accepted as just as necessary in scholarly book publishing as in scholarly journal publishing in the fields where monographs and book
chapters play an important role in research communication, namely the social sciences and humanities (Giménez-Toledo et al., 2015).

Peer review has also increasingly become a delineation criterion in national or regional systems providing data for the measurement or assessment of scholarly publishing in books. This is the case in Norway (Sivertsen and Larsen, 2012; NSD, 2017) Finland (Pölönen & Ruth, 2015), Denmark (Ingwersen & Larsen, 2014), Flanders (Verleysen & Engels, 2013) and Spain (Giménez-Toledo et al., 2015). Either at the level of the publisher (e.g. Norway’s Cristin system) or at the level of the individual book (e.g. Flanders’ VABB-SHW and Spain’s CNEAI (BOE, 2016), peer review by external experts in the field is required. This requirement may even be expressed in the definition of scholarly publishing on which the system is based. The requirement of peer review may also be present in information systems and international databases that are not directly linked to research assessment, such as DOAB, Scopus, Book Citation Index, Kriterium or Open Edition.

Nevertheless, peer review practices may be even more diverse in scholarly publishing than in scholarly journal publishing. Most scholarly publishers are hybrid in the sense that books representing original research and oriented at scholarly publications are only a part of their publishing profile. Books for students and wider audiences can be just as important for the publisher, and not only scholarly criteria, but also commercial criteria, may play a role in the decision what to publish. Not all publishers that frequently review manuscripts from academia apply the same rigorous selection processes. ‘Peer review’ can be understood in several different forms in the case of scholarly books.

This work has the objective of analyzing how the ‘peer reviewed’ criterion is actually being applied when categorizing or classifying scholarly publishers in national or regional systems for research assessment and bibliometric measurement, mainly in the case of Social Sciences and Humanities. Exploring and comparing the different practices carried out internationally can provide mutual learning and harmonized guidelines as a next step. A further step could be an integrated international information system that includes also scholarly book with standardized information on publishers. Two existing projects, the so-called VIRTA solution originating from Finland, and ERIH PLUS, so far covering only journals, have such international integration as one of the aims.

The existence of national information systems is crucial since it allows better understanding and reflecting on the publishing diversity of a country – both in terms of number of publishers, scholarly fields or publishing languages. But it should be possible to compare those categorizations of scholarly books across Europe. For that purpose, comparing how peer review is defined, and how this definition delineates the data in practice, is necessary. We also need to understand how the applied definition influences bibliometric measurement.
Scholarly book publishing in the humanities and social sciences takes place and is important both at an international and at a national level. The project may result in mutual learning between to two levels, which can be useful for the publishers as well.

In Finland and Belgium, a specific label for publishers and books – respectively – has been introduced for the aid of delineating peer reviewed scholarly publishing. Both countries, as well as Denmark and Norway, monitor lists of approved scholarly publishers with peer review procedures as a delimiting criterion. In Spain, the review processes described in SPI (Scholarly Publishers Indicators) is used as a reference by evaluation agencies. This set of countries has been selected because in all cases they count with lists of publishers and detailed specifications concerning how peer review is considered in evaluation processes, although it is among the aims of this line of research to comparatively study peer review in the evaluation systems of other countries. Also, supra-national evaluation systems might count with specifications concerning the peer review of books, but these have not been the focus of this research since the study of national evaluation processes offers a more detailed and specific ground for further convergence or integration.

Methodology

In this first approach to the commonalities on the treatment of scholarly books’ peer review, the documentation and information publicly available at the country level in each case are analyzed and summarized. The methodology primarily consists on the search, download and analysis of documents available in the websites of the four previously mentioned systems, or in the legislation concerning the databases. The set of fields reflected in the juxtaposition table have been chosen according to the frequency and relevance of their appearance in the documentation analyzed.

Results

The following table (1) reflects how peer review is defined in each country, concerning specifically peer review practices for scholarly books.

The Norwegian and Danish systems apply a specific definition of peer review, while the Finnish and Flemish systems apply an explicit set of conditions that series and individual titles (only individual titles in the case of Flanders’ GPRC, Guaranteed Peer Review Content label) must meet in order to obtain the label attributing the publisher or the title a controlled and certified peer review process. In Spain, the criteria which publishers must meet for entering the evaluation system are set in the evaluation regulations and SPI (Scholarly Publishers Indicators) provides the information on peer review which is afterwards considered by evaluation agencies. In the case of Flanders, since the general set of conditions is different from the specifications for the GPRC label, an entry for each type of information has been added to the table below.
<table>
<thead>
<tr>
<th>Country</th>
<th>Peer review must occur before publication</th>
<th>Anonymity of peer review</th>
<th>Reviewers autonomy or reviewers independence</th>
<th>Number of reviewers</th>
<th>Proofing documentation</th>
<th>Key issues in the review process</th>
<th>Source of definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norway</td>
<td>Yes, by definition</td>
<td>Not necessarily (either way)</td>
<td>At least one must be ‘without ties to the publisher or the author’</td>
<td>Not specified</td>
<td>Requisite for inclusion in lists. Not in definition.</td>
<td>Originality and quality</td>
<td><a href="https://dbh.nsd.uib.no/publiseringskanaler/OmKriterier">https://dbh.nsd.uib.no/publiseringskanaler/OmKriterier</a></td>
</tr>
<tr>
<td>Finland (Definitio n operative at the label level)</td>
<td>Implicit</td>
<td>Not necessarily (either way)</td>
<td>Independent from author and editorial board</td>
<td>At least two</td>
<td>Requisite for inclusion and for label. Detailed set of requirements</td>
<td>Comprehensiveness of the material and the mastering of the theoretical framework, reliability and accuracy of the research</td>
<td><a href="https://www.tsv.fi/en/services/label-for-peer-reviewed-scholarly-publications/requi">https://www.tsv.fi/en/services/label-for-peer-reviewed-scholarly-publications/requi</a> rements-for-use</td>
</tr>
<tr>
<td>Denmark</td>
<td>Yes, by definition</td>
<td>Not necessarily</td>
<td>At least one must be external to the publisher/institution</td>
<td>Not specified</td>
<td>In order to count in the BFI scheme, it is assumed that documentation proving peer review can be supplied</td>
<td>Scientific quality and originality</td>
<td><a href="http://ufm.dk/forskningsog-innovation/statistik-og-analyser/den-bibliometriske-forskningsindikator/bfis-regler">http://ufm.dk/forskningsog-innovation/statistik-og-analyser/den-bibliometriske-forskningsindikator/bfis-regler</a></td>
</tr>
<tr>
<td>Flanders (General definition )</td>
<td>Yes, by definition</td>
<td>Not necessarily</td>
<td>At least one must be independent from research team and authors. Peer review cannot be organized by the author</td>
<td>Not specified but detailed in plural</td>
<td>Not specified</td>
<td>Not specified.</td>
<td><a href="http://data-onderwijs.vlaanderen.be/edulex/document.aspx?docid=14492#2">http://data-onderwijs.vlaanderen.be/edulex/document.aspx?docid=14492#2</a></td>
</tr>
</tbody>
</table>
Transparency concerning peer review practices: The inclusion of an approved scholarly publisher in the Norwegian and Finnish information systems requires the existence of publicly available information provided by the publishers concerning their peer review practices. In the case of the Flemish system, there is no formal requirement for the publishers to have publicly available information concerning their peer review processes. However, the Authoritative Panel that controls what data enters VABB-SSHW, make a decision on the use of peer review by publishers in the system. In the BFI Danish scheme, there is also no formal requirement of publicity concerning the use of peer review by publishers, but any publication channel in the system is expected to be able to provide proofs in that sense. In the Spanish case, the requirement is implicit in the evaluation regulations concerning the use data from SPI with regards to scholarly publishers’ peer review procedures.

Preliminary conclusions

The analysis of the definitions and uses of peer review on scholarly books in the different countries allows for the following preliminary conclusions:

1. Peer review is a delineating criterion for the evaluation of scholarly outputs in the form of books, monographs and book series. It is present in all four systems and an explicit prerequisite for the inclusion of a publication channel as suitable for further evaluation.

2. There is a diversity of approaches to defining peer review and applying it in the evaluation process: from the specific definitions and requirements in the case of the Finnish and Flemish labels to the use of existing information on scholarly publishers’ peer review practices by evaluation agencies in the case of Spain.
3. Peer review of scholarly books needs further research to find out which features should define the due criteria for peer review processes, which weight would be appropriate to attribute to peer reviewed contributions and which level of analysis is optimal in terms of cost and precision (from publisher to individual titles).

4. The possible influence of commercial criteria in peer reviewed scholarly publishing should also be studied further.

5. Lastly, further research is needed to lay the basis for international data integration and comparable measurements.

We also expected to see how specific publishers (international or national) are classified, in terms of peer review practices, in the different information systems. The possibility of creating a shared list for all of the systems will also be investigated. Another possible line of further research would include the study of peer review and evaluation practices applied to book series such as the Finnish quality label and, in the case of Spain, the UNE (Union of University Presses) quality label for series (Academic Publishing Quality Label / APQ: CEA-APQ, 2017). The above mentioned Handbook of Best Practices for Peer Review (AAUP, 2016) may serve as a kind of yardstick for the comparison of systems.

Acknowledgements:

This work has been supported by Research Project (CSO2015-63693-P), funded by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund. This article is based upon work from COST Action (CA15137), supported by COST (European Cooperation in Science and Technology).

References:


NSD (NORWEGIAN REGISTER FOR SCIENTIFIC JOURNALS, SERIES AND PUBLISHERS), 2017. Procedures for processing new submissions. Available at: https://dbh.nsd.uib.no/publiseringskanaler/OmProsedyrer


Andrey Lovakov¹  Elena Agadullina²

¹lovakov@hse.ru
National Research University Higher School of Economics, Moscow (Russia)

²eagadullina@hse.ru
National Research University Higher School of Economics, Moscow (Russia)

Abstract
During several decades Soviet academic psychology community was isolated from the West, but after the collapse of the Soviet Union each of the 15 countries went its own way in economic, social, and scientific development. The paper analyses publications from post-Soviet countries in psychological journals in 1992–2016, i.e. 25 years after the collapse of the Soviet Union. Results show that 15 post-Soviet countries have produced in sum less than one percent from the world output in psychological journals. There is a huge diversity in the number of papers between 15 post-Soviet countries. Russia, Estonia, and Lithuania are the leaders among them. Authors of the more than 90% of all post-Soviet countries’ papers are affiliated with these three countries. The most intensive collaboration is between Russia, Estonia, Lithuania, and Georgia and between three Baltic countries. Post-Soviet countries also differ in publication patterns.

Conference Topic
Country-level studies

Keywords
Post-Soviet countries; psychological journals; bibliometric analysis

Introduction
Internationalization and global communication and collaboration are the key attributes of the contemporary science. To be a part of international community means publishing papers in international journals and in English language because English is a lingua franca for science today. This is true for the majority scientific fields including psychology. Despite the fact that psychological academic community was relatively open for international communication in the early Soviet period (Yasnitsky, 2010, 2011) and a number of Soviet psychologists and behavioural researchers were recognized by the international community (Aleksandrova-Howell, Abramson, & Craig, 2012; Haggblom et al., 2002) a long time Soviet academic psychology community was isolated from the West. Ideological pressure in social sciences and the Iron Curtain leaded to almost full isolation from western scientific community. After the collapse of the Soviet Union each of the 15 countries, which were parts of the Soviet Union, went their own ways in economic, social, and scientific development. There were a number of administrative reforms and initiatives aimed at improvement and development of scientific research systems in post-Soviet countries during the last decades (Abbott & Schiermeier, 2014; Kristapsons, Martinson, & Dabyte, 2003; Raudla, Karo, Valdmaa, & Kattel, 2015; Schiermeier, 2010; Turko, Bakhiturin, Bagan, Poloskov, & Gudym, 2016).

There are several bibliometric studies of scientific developments in the former Soviet countries (Allik, 2008, 2015; Fiala & Willett, 2015; Gzoyan et al., 2015; Kozak, Bornmann, & Leydesdorff, 2015; Markusova, Ivanov, & Varshavskii, 2009; Markusova, Jansz, Libkind, & Varshavsky, 2009; Zavadskas, Kirvaitis, & Dagiene, 2011). But these studies usually consider only several countries and analyze publications from all fields as a whole

125
output or focus on natural/computer sciences. While social and behavioural sciences were given little attention. The main goal of this paper is to provide a bibliometric analysis of publications from 15 post-Soviet countries in psychological journals. This study tries to answer the following research questions. What is the contribution of the former Soviet countries into the world psychological science? How do the academic communities of psychologists in these countries collaborate with each other and how they are integrated into the international community?

Data and Methods
Data were obtained from Web of Science (Clarivate Analytics, ex-Thomson Reuters), the online version of the SCI-EXPANDED and SSCI, on 20th February 2017. An advanced search by fields Web of Science Category (WC), Country (CU), and Year Published (PY) was conducted. We searched articles with at least one author affiliated with one of 15 post-Soviet countries in journals belonging to at least one psychology category: Psychology; Psychology, Applied; Psychology, Biological; Psychology, Clinical; Psychology, Developmental; Psychology, Educational; Psychology, Experimental; Psychology, Mathematical; Psychology, Multidisciplinary; Psychology, Social. The retrieved results were limited by document types (‘article’ or ‘review’), language (English), and year of publication (1992–2016). The full search string is in Appendix. With this search strategy 1950 articles were found. All available information about these 1950 articles was downloaded. For the data preparation, analysis and visualization we used R, a programming language and environment for statistical computing (R Core Team, 2016). All results described in this article, were extracted and calculated by the authors' own R script. Additional packages were used: ggplot2 (Wickham, 2009), readxl (Wickham, 2016), stringr (Wickham, 2017), stringi (Gagolewski & Tartanus, 2016), reshape2 (Wickham, 2007).

Analysis of the international collaborations was conducted by the freely available computer program VOSviewer (version 1.6.5) (van Eck & Waltman, 2010). Data on international collaborations were saved from VOSviewer and visualized by the igraph package (Csardi & Nepusz, 2006).

Results and Discussion
In sum 15 post-Soviet countries produced 1950 papers indexed in SCI-EXPANDED and SSCI. It is 0.34% from all 581,425 articles in psychological journals published by all countries. The share of the papers in psychological journals from these countries increased from 0.01% to 0.61% during the considered interval (Figure 1). Though it is still less than one percent from the world output, there is a growth in the contribution of post-Soviet countries to the world output in psychological journals. Spearman's rank correlation between publication year and the percentage of the post-Soviet countries’ papers from the world output in psychological journals is 0.78 (p < 0.001). It is interesting that this growth is accompanied by a decrease of the share of all papers from post-Soviet countries in the world journals (Figure 1). Spearman's rank correlation between publication year and the percentage of the post-Soviet countries’ papers from the world output in all journals is –0.45 (p = 0.025).
Figure 1. The share of the post-Soviet countries in the world output in psychological journals and in all journals (authors’ calculations based on SCI-EXPANDED and SSCI data).

Table 1. The number of papers from post-Soviet countries in psychological journals (1992–2016).

<table>
<thead>
<tr>
<th>Country</th>
<th>Total number of papers</th>
<th>Percentage from all post-Soviet countries' papers</th>
<th>Total number of citations</th>
<th>The number of citations per paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armenia</td>
<td>11</td>
<td>0.6</td>
<td>142</td>
<td>12.91</td>
</tr>
<tr>
<td>Azerbaijan</td>
<td>6</td>
<td>0.3</td>
<td>67</td>
<td>11.17</td>
</tr>
<tr>
<td>Byelarus</td>
<td>25</td>
<td>1.3</td>
<td>202</td>
<td>8.08</td>
</tr>
<tr>
<td>Estonia</td>
<td>537</td>
<td>27.5</td>
<td>8643</td>
<td>16.09</td>
</tr>
<tr>
<td>Georgia</td>
<td>62</td>
<td>3.2</td>
<td>1004</td>
<td>16.19</td>
</tr>
<tr>
<td>Kazakhstan</td>
<td>17</td>
<td>0.9</td>
<td>73</td>
<td>4.29</td>
</tr>
<tr>
<td>Kyrgyzstan</td>
<td>7</td>
<td>0.4</td>
<td>24</td>
<td>3.43</td>
</tr>
<tr>
<td>Latvia</td>
<td>38</td>
<td>1.9</td>
<td>863</td>
<td>22.71</td>
</tr>
<tr>
<td>Lithuania</td>
<td>146</td>
<td>7.5</td>
<td>1810</td>
<td>12.4</td>
</tr>
<tr>
<td>Moldova</td>
<td>3</td>
<td>0.2</td>
<td>92</td>
<td>30.67</td>
</tr>
<tr>
<td>Russia</td>
<td>1099</td>
<td>56.4</td>
<td>14613</td>
<td>13.3</td>
</tr>
<tr>
<td>Tajikistan</td>
<td>2</td>
<td>0.1</td>
<td>11</td>
<td>5.5</td>
</tr>
<tr>
<td>Turkmenistan</td>
<td>2</td>
<td>0.1</td>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>Ukraine</td>
<td>72</td>
<td>3.7</td>
<td>839</td>
<td>11.65</td>
</tr>
<tr>
<td>Uzbekistan</td>
<td>12</td>
<td>0.6</td>
<td>67</td>
<td>5.58</td>
</tr>
</tbody>
</table>

Three countries have the largest number of papers: Russia, Estonia, and Lithuania (Table 1). Authors of the more than 90% of all post-Soviet countries’ papers are affiliated with these three countries. Figure 2 shows distribution of papers from post-Soviet countries in psychological journals between 1992 and 2016. The analysis shows that Russia and Estonia are the leaders during all this period. Before 2011–2012 there was a slow linear growth of the number of papers in both countries, but after that there is a sharp rise in the number of papers from these countries. Perhaps, this rise can be explained by several policy decisions and initiatives in Russian and Estonian academic systems, which have affected all scientific fields.
including psychology. In Russia these are grants to support scientific research projects implemented by the world’s leading scientists at Russian institutions of higher learning (in Russia, so called “megagrants”) and Russian Academic Excellence Project “5-100” aiming to maximize the competitive position of a group of leading Russian universities in the global research and education market (Schiermeier, 2010). Previous research showed that every fifth Russian article published in psychological journals in 2015 was supported by megagrants (Lovakov, in press). In Estonia this is primarily project-based funding of research (more than 80% of research funding is project-based) (Allik, 2015; Raudla et al., 2015).

The number of papers from Georgia has also increased in recent years. Eight post-Soviet countries (Armenia, Azerbaijan, Kazakhstan, Kyrgyzstan, Moldova, Tajikistan, Turkmenistan, and Uzbekistan) have less than 20 papers in psychological journals during the last 25 years. This means less than one paper per year.

![Figure 2. The number of papers in psychological journals from post-Soviet countries (only countries with 20 and more papers).](image)

Psychology is a complex discipline, this is reflected by several special psychological categories in SSCI. Table 2 shows the distribution of papers across nine psychological categories. Papers from the post-Soviet countries are presented in journals from all categories. But most of the papers are concentrated in journals from four categories: multidisciplinary, experimental, social, and biological. The papers in journals from other categories are poorly represented. The distribution of papers between categories remains unchanged during all the analyzed period. Due to the limitations of the journals classification into WoS categories it is useful to look at the papers' distribution among the journals itself. Table 3 shows top 10 titles with the highest number of post-Soviet publications for three periods. In 1992–2000 papers from biological and cognitive science journals dominated among the post-Soviet countries’ papers. But after 2000 papers from social and personality journals also emerged. The authors from the post-Soviet countries still publish rarely in the prestigious journals on clinical psychology.
Table 2. Distribution (percent) of papers from post-Soviet countries in psychological journals across WoS categories by three time intervals.

<table>
<thead>
<tr>
<th>WoS category</th>
<th>Percentage of papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychology, Multidisciplinary</td>
<td>21.6</td>
</tr>
<tr>
<td>Psychology, Experimental</td>
<td>23.4</td>
</tr>
<tr>
<td>Psychology, Social</td>
<td>12.2</td>
</tr>
<tr>
<td>Psychology, Biological</td>
<td>22.8</td>
</tr>
<tr>
<td>Psychology, Clinical</td>
<td>10.9</td>
</tr>
<tr>
<td>Psychology, Developmental</td>
<td>7.6</td>
</tr>
<tr>
<td>Psychology, Educational</td>
<td>4.9</td>
</tr>
<tr>
<td>Psychology, Mathematical</td>
<td>2.4</td>
</tr>
<tr>
<td>Psychology, Applied</td>
<td>5.5</td>
</tr>
</tbody>
</table>

Table 3. Ten titles with the highest number of post-Soviet papers for three periods.

<table>
<thead>
<tr>
<th>Journal</th>
<th>The number of papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992–2000</td>
<td></td>
</tr>
<tr>
<td>International Journal of Psychophysiology</td>
<td>26</td>
</tr>
<tr>
<td>Behavior Genetics</td>
<td>16</td>
</tr>
<tr>
<td>Physiology &amp; Behavior</td>
<td>16</td>
</tr>
<tr>
<td>International Journal of Mental Health</td>
<td>11</td>
</tr>
<tr>
<td>Integrative physiological and behavioral science</td>
<td>10</td>
</tr>
<tr>
<td>Aggressive behavior</td>
<td>8</td>
</tr>
<tr>
<td>Personality and Individual Differences</td>
<td>8</td>
</tr>
<tr>
<td>Ethology</td>
<td>6</td>
</tr>
<tr>
<td>Behavioral and Brain Sciences</td>
<td>5</td>
</tr>
<tr>
<td>Perception</td>
<td>5</td>
</tr>
<tr>
<td>2001–2010</td>
<td></td>
</tr>
<tr>
<td>International Journal of Psychophysiology</td>
<td>31</td>
</tr>
<tr>
<td>Personality and Individual Differences</td>
<td>28</td>
</tr>
<tr>
<td>Spanish Journal of Psychology</td>
<td>25</td>
</tr>
<tr>
<td>Perceptual and Motor Skills</td>
<td>22</td>
</tr>
<tr>
<td>Physiology &amp; Behavior</td>
<td>20</td>
</tr>
<tr>
<td>Journal of Cross-Cultural Psychology</td>
<td>18</td>
</tr>
<tr>
<td>AIDS Care. Psychological and Socio-medical Aspects of AIDS/HIV</td>
<td>14</td>
</tr>
<tr>
<td>Perception</td>
<td>13</td>
</tr>
<tr>
<td>European Journal of Personality</td>
<td>12</td>
</tr>
<tr>
<td>Culture &amp; Psychology</td>
<td>11</td>
</tr>
<tr>
<td>2011–2016</td>
<td></td>
</tr>
<tr>
<td>Frontiers in Psychology</td>
<td>32</td>
</tr>
<tr>
<td>Journal of Cross-Cultural Psychology</td>
<td>28</td>
</tr>
<tr>
<td>Intelligence</td>
<td>25</td>
</tr>
<tr>
<td>Learning and Individual Differences</td>
<td>25</td>
</tr>
<tr>
<td>Personality and Individual Differences</td>
<td>25</td>
</tr>
<tr>
<td>International Journal of Psychophysiology</td>
<td>24</td>
</tr>
<tr>
<td>Frontiers in Human Neuroscience</td>
<td>22</td>
</tr>
<tr>
<td>Physiology &amp; Behavior</td>
<td>21</td>
</tr>
<tr>
<td>Perception</td>
<td>20</td>
</tr>
<tr>
<td>AIDS Care. Psychological and Socio-medical Aspects of AIDS/HIV</td>
<td>18</td>
</tr>
</tbody>
</table>
The probable explanation may be that social sciences were under ideological pressure to a greater extent than natural and life sciences during the Soviet period. Therefore, after the collapse of the Soviet Union the social part of psychological science was more isolated and backward than the natural one. For this reason, in the 1990s, researchers from the social part of psychological science found it more difficult to produce research and publish papers that may become competitive at the international level.

Figure 3 shows the percentage of international papers in all countries together, as well as separately for Russia and Estonia. In the first period (before 2000), the share of international articles among articles of all countries in psychological journals ranges from 0% to 50%, and in the second period the share varies around 60%. Whereas the collaborations between post-Soviet countries were only episodic and only between several countries in the 90s. It means that right after gaining independence former members of Soviet Union were publishing papers by themselves or in collaboration with some other countries, but not in collaboration with post-Soviet countries. It is possible that at that time it was easier or more profitable to collaborate with other countries, because most of post-Soviet countries were in economic and political crisis.

On the regular basis papers in co-authorship with researchers from other post-Soviet countries emerged only in the 2000s. The greatest diversity in the co-authorship of researchers from post-Soviet countries is observed in the period from 2001 to 2010. The most intensive collaboration was between Russia, Estonia, Lithuania, and Georgia. Researchers from three Baltic countries also collaborate between each other. Even during the whole period from 1992 to 2016 not all post-Soviet countries collaborate with each other in psychology. Three of the fifteen countries (Kazakhstan, Tajikistan, and Turkmenistan) do not have any papers published in co-authorship with other post-Soviet countries (Fig. 4D). The small number of papers in psychological journals from these countries is the main reason.

![Figure 3. Percentage of papers with international collaboration in psychological journals (authors’ calculations based on SCI-EXPANDED and SSCI data).](image-url)
Country-based analysis

Further bibliometric analysis is based on aggregate indicators, which are informative only for countries with significant enough number of papers. Due to this, in further analysis we will consider papers from only seven post-Soviet countries (Byelarus, Estonia, Georgia, Latvia, Lithuania, Russia, and Ukraine) with more than 20 papers.

Russian papers concentrate mostly in journals from Multidisciplinary, Experimental, and Biological categories (Table 4). More than half of Estonian papers distribute between journals from Multidisciplinary, Experimental, and Social categories. Lithuanian papers are published mostly in journals from Multidisciplinary, Experimental, Developmental, and Clinical WoS categories. Papers from all post-Soviet countries published in journals from Educational, Mathematical, and Applied WoS categories are relatively rare.

Figure 4. Co-authorship networks of post-Soviet countries for the whole period from 1992 to 2016 and for three periods (1992–2000, 2001–2010, 2011–2016, and 1992–2016) separately. The more thick the edge, the closer the collaboration is among the researchers from the countries.
Table 4. Distribution (percent) of papers from post-Soviet countries in psychological journals across WoS categories.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Byelarus</td>
<td>24.0</td>
<td>8.0</td>
<td>16.0</td>
<td>20</td>
<td>12.0</td>
<td>8.0</td>
<td>4.0</td>
<td>4.0</td>
<td>0</td>
</tr>
<tr>
<td>Estonia</td>
<td>20.9</td>
<td>22.9</td>
<td>23.8</td>
<td>8.0</td>
<td>7.3</td>
<td>10.1</td>
<td>8.2</td>
<td>0.2</td>
<td>3.5</td>
</tr>
<tr>
<td>Georgia</td>
<td>12.9</td>
<td>8.1</td>
<td>21.0</td>
<td>1.6</td>
<td>30.6</td>
<td>11.3</td>
<td>1.6</td>
<td>1.6</td>
<td>9.7</td>
</tr>
<tr>
<td>Latvia</td>
<td>31.6</td>
<td>10.5</td>
<td>15.8</td>
<td>18.4</td>
<td>7.9</td>
<td>5.3</td>
<td>7.9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lithuania</td>
<td>26.0</td>
<td>18.5</td>
<td>14.4</td>
<td>2.7</td>
<td>15.1</td>
<td>16.4</td>
<td>4.1</td>
<td>0.7</td>
<td>7.5</td>
</tr>
<tr>
<td>Russia</td>
<td>22.0</td>
<td>23.8</td>
<td>12.7</td>
<td>18.2</td>
<td>9.3</td>
<td>11.0</td>
<td>5.6</td>
<td>2.7</td>
<td>3.2</td>
</tr>
<tr>
<td>Ukraine</td>
<td>27.8</td>
<td>16.7</td>
<td>8.3</td>
<td>12.5</td>
<td>20.8</td>
<td>9.7</td>
<td>0</td>
<td>2.8</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Note: The sum by rows may be greater than 100% because some journals belong to two or more categories.

Table 5. Characteristics of papers from post-Soviet countries in psychological journals.

<table>
<thead>
<tr>
<th>Country</th>
<th>Number of authors per paper</th>
<th>Number of countries per paper</th>
<th>Percentage of papers with authors from one country</th>
<th>Percentage of first-authored papers</th>
<th>Percentage of papers in top journals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Min</td>
<td>Max</td>
<td>Mean</td>
<td>Min</td>
</tr>
<tr>
<td>Byelarus</td>
<td>5.68</td>
<td>1</td>
<td>38</td>
<td>2.76</td>
<td>1</td>
</tr>
<tr>
<td>Estonia</td>
<td>7.29</td>
<td>1</td>
<td>131</td>
<td>3.47</td>
<td>1</td>
</tr>
<tr>
<td>Georgia</td>
<td>14.95</td>
<td>1</td>
<td>143</td>
<td>7.45</td>
<td>1</td>
</tr>
<tr>
<td>Latvia</td>
<td>19.32</td>
<td>2</td>
<td>131</td>
<td>7.26</td>
<td>1</td>
</tr>
<tr>
<td>Lithuania</td>
<td>10.98</td>
<td>1</td>
<td>131</td>
<td>5.63</td>
<td>1</td>
</tr>
<tr>
<td>Russia</td>
<td>5.16</td>
<td>1</td>
<td>143</td>
<td>2.63</td>
<td>1</td>
</tr>
<tr>
<td>Ukraine</td>
<td>5.83</td>
<td>1</td>
<td>67</td>
<td>3.11</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: Top journals are those from the first or second quartiles at least in one psychological WoS category.

Table 5 shows characteristics of papers from seven post-Soviet countries. Based on these data it is possible to identify several publication patterns among countries. Georgia, Latvia, and Lithuania have the highest average number of authors per paper, the highest average number of countries per paper. It means that these countries have published a lot of papers in big international collaborations. At the same time, Georgia published only 24% of papers where an author from this country is a first author, whereas Latvia and Lithuania have nearly a half of such papers. In psychology the order of the authors should reflect their contribution to the article (Publication Manual of the American Psychological Association, 2010). The first author or reprint author usually has the largest contribution to the paper. Thus, Lithuania published papers both independently (34%) and in international co-authorship including big collaborations. And Georgia and Latvia published papers mostly in big international co-operation. Russia and Ukraine also have low percentage of papers where authors from these countries are the first authors and where all the authors are affiliated with the same country. Thus, authors from these countries mostly published papers in international collaboration where they are not the main authors. In contrast, Estonian and Byelorussian authors published half of their papers without international collaboration, and in more than half of their papers they are the main authors.

Conclusion

We have analysed publications from post-Soviet countries in psychological journals in 1992–2016, i.e. during 25 years after the collapse of the Soviet Union.
The main conclusion of our research is that the output of psychology papers in the leading international journals is rather poor after the dissolution of the USSR, but the ways of development of psychological science in the countries which formed the Soviet Union differ significantly. Bibliometrics shows this very well.

In sum, the former Soviet republics have produced less than one percent of the world output in psychological journals. However the growth of post-Soviet countries’ share in the world psychology is observed.

There is a huge diversity in the number of psychology papers between 15 post-Soviet countries. Russia, Estonia, and Lithuania are the leaders among them. Countries also differ in their publication patterns. Authors from one group of countries (Georgia and Latvia) published papers mostly in big international collaborations, while authors from another group (Estonia and Byelorus) published a lot of papers without international collaboration or, if in collaboration, then their author was usually the main author of the text who has the most important contribution to the paper.

More detailed investigation on how the development of academic psychology has become so different in various post-Soviet republics needs an analysis on the institutional level, this will become an interesting topic for a further research.

Acknowledgments
The paper was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE), and supported within the framework of a subsidy by the Russian Academic Excellence Project ‘5-100’.

References


**Appendix**

Search string

(WC=""PSYCHOLOGY" OR "PSYCHOLOGY, APPLIED" OR "PSYCHOLOGY, BIOLOGICAL" OR "PSYCHOLOGY, CLINICAL" OR "PSYCHOLOGY, DEVELOPMENTAL" OR "PSYCHOLOGY, EDUCATIONAL" OR "PSYCHOLOGY, EXPERIMENTAL" OR "PSYCHOLOGY, MATHEMATICAL" OR "PSYCHOLOGY, MULTIDISCIPLINARY" OR "PSYCHOLOGY, SOCIAL") AND CU=(armenia OR azerbaijan OR byelarus OR estonia OR georgia OR kazakhstan OR kyrgyzstan OR latvia OR lithuania OR moldova OR russia OR tajikistan OR turkmenistan OR ukraine OR uzbekistan)
AND PY=1992-2016) AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article OR Review)

Indexes=SCI-EXPANDED, SSCI Timespan=All years
Global overview of unmanned aerial vehicles research: country-level and organisation-level bibliometric analysis

Maxim Kotsemir¹

¹ mkotsemir@hse.ru
National Research University Higher School of Economics, Institute for Statistical Studies and Economics of Knowledge, Moscow (Russia)

Abstract

This study proposes the global bibliometric overview of unmanned aerial vehicles (UAVs) research in Scopus database in 1985 – 2015. This study detects key countries in this field of research as well as the major centers of excellence (organisations) in UAV research. We analyse publication activity of leading countries and organisations as well as the level of citation of their UAV publications. Special section is devoted to the analysis of cross-country collaboration links. For plotting the map of international collaboration in UAV research, VOSviewer software was used.

Keywords: Unmanned aerial vehicles; UAV; aviation; publication activity; Scopus; VOSviewer; bibliometric analysis; international collaboration; citation analysis; scientometrics.

Conference Topic:
Journals, databases and electronic publications
Country-level studies
Mapping and visualization
Introduction

Unmanned aerial vehicles (UAVs) become a new special sector on aviation industry. The application of unmanned aerial vehicles is widespread: from giant global Hawk to micro aerial vehicles used in precise agriculture. Unmanned aerial vehicles research is related with many different fields of science. First there are spheres of research directly related with aviation science. UAV research posts new questions in fields of aerodynamics. The key questions here is how to measure, control and improve the aerodynamic characteristics on UAVs of untraditional forms: flying wing, spherical UAVs, UAVs with flexible and flapping wings, insect-like UAVs, quadricopters, etc. Principally new avionics is needed for UAVs. In avionics for UAVs the key role plays artificial intelligence controlled avionics or avionics system that are remotely controlled by human. New paradigm of flight management systems (FMS) is needed. With the fast development of unmanned aerial vehicles industry the new challenges for aviation safety (primarily related with antiterrorist control) arise. UAVs also change the essence of military aviation.

It should be said that there was only several examples of bibliometric analysis of research in the field of aviation and aerospace industry:

- Studies devoted to the bibliometric analysis of aerospace and aeronautical journals (Meskoob and Tanbakouei, 2012; Nageswara Rao et al. 2014; Aswathy and Pal, 2015);
- General bibliometric analysis of specific subfields of aerospace science (Ganguli, 2008; Evans, et al. 2008, 2009; Rezadad & Maghami, 2014; Rojas-Sola and Aguilera-Garcia, 2014);
- Technological mining in specific subfields of aerospace science based on bibliometric and patent analysis (Nakamura et al., 2010, 2012; Xu and Hua, 2014; Li and Guo; 2015).
- Analysis of publication activity in some organizations (and its divisions) related with aerospace research (Stephens, 2013; Nakamura, Kajikawa, and Suzuki; 2014).

This paper is among the first studies that provide the global overview of publication activity in field of unmanned aerial vehicles. This study detects key countries and key centers of excellence in this field.

1. Methodological issues

To run the bibliometric analysis on unmanned aerial vehicles (UAV) it is needed first to determine the field of search for these publications. Two greatest multisubject science citation databases that are recognized by the international scientific community and have well-developed interface for a bibliometric analysis are Scopus and Web of Science. These databases among other possibilities for bibliometric analysis allow the user to search any term (as well as combination of terms) in title, abstract and keywords of publications. In this paper, Scopus database was used since Web of Science database does not have the system of unique author identifiers and unique organization identifiers. Therefore, the possibilities for the analysis of publication activity of individual organizations and authors in Web of Science are seriously

1 The various aspects of research on the aerodynamic characteristics of unmanned aerial vehicles are studied in e.g. Ansari et al., Curet et al., 2013.
2 Some examples of research on avionics in application to unmanned aerial vehicles are e.g. Naruoka et al. 2009; Sudha Rani and Ramadoss, 2015.
3 Examples of studies on flight management systems within UAV research are: Koo et al., 1999; Pasaoglu et al., 2016.
4 Discussion on problems of hacking and hostile takeover of UAV control can be found in: Kim et al., 2012; Dulo, 2015.
5 Examples of studies on military application of UAVs are: Tozer at al., 2000; Gowtham and Gnanasundari, 2015.
constrained in contrast to Scopus⁶. Moreover, in Scopus the number of publications on unmanned aerial vehicles in different periods is 1.5 – 3.5 times higher than in Web of Science.

To detect publications on UAVs the search in title or abstract or keywords of publications was done the following terms (and all forms of them): unmanned aerial vehicle; unmanned airborne vehicle; unmanned flying vehicle; pilotless aerial vehicle; pilotless flying vehicle; unmanned aircraft; pilotless aircraft; unmanned helicopter; pilotless helicopter; unmanned copter; pilotless copter; unmanned plane; pilotless plane; unmanned drone; pilotless drone; air drone; aerial drone; flying drone; quadrocopter; quadricopter; micro aerial vehicle, micro air vehicle; micro flying vehicle; micro aircraft system. Such abbreviations like UAV, UAVs, MAV MUAV etc. were not included into the search query in order not to take irrelevant words like “User Antenna View”; “Mars Ascent Vehicle”, “Modular Acoustic Velocity Sensor” etc. We should understand: when we search for the word “unmanned Ariel vehicle” in title, abstract or keywords of the publication we take into our analysis not only publications devoted directly to unmanned aerial vehicles but also publications where the word combination "unmanned aerial vehicle" is simply mentioned. On the other hand mentioning UAV-related term(-s) in abstract or keywords of a given publication means that this publication is at least indirectly related with unmanned aerial vehicles. Moreover, the search only in the title of publications is a very narrow approach. In many cases author may not mention the word "unmanned aerial vehicles" in the title when the publication is devoted to use of some material or some technologies in unmanned aerial vehicles or new ways of application of UAVs. Therefore, in our case we treat all publications where the above mentioned UAV search terms are mentioned in title or keywords or abstract as publications related with unmanned aerial vehicles. All types of documents and all types of sources were included in the analysis.

2. Country-level analysis of publication activity in UAV studies

Since more or less stable dynamics of publications in Scopus database starts since 1985, we analyse the dynamics of publication activity on UAVs for the period of 1985 – 2015. There were several periods of rapid growth of global number of UAV publications in Scopus: 2001 – 2007 and 2012 – 2013.

Figure 1. Evolution of unmanned aerial vehicles research in Scopus in 1985 – 2015.

Notes: All types of documents and sources indexed in Scopus are included in the analysis. The analysis was done in 2nd decade of January 2016.

⁶ See more on advantages and limitations of Scopus vs. Web of Science in e.g. Meho and Yang, 2006; Falagas et al, 2008; Archambault et al., 2009; Vieira and Gomes, 2009; Shashnov and Kotsemir, 2015.
The United States is the major player in global UAV studies (Figure 2). For 1985 – 2015 USA contributed 35.8% of global come of publications on UAVs. The other big player is China whose contribution for 1985 – 2015 was 20.4%. The rocket growth of publication activity in UAVs in China started in 2005. The share of China in global volume of publications on UAVs has increased from 4.9% in 2002 to 26.5% in 2015\(^7\). The share of the United States on the contrary decreased from 68.8% in 2000 to 23.7% in 2015 (Table A.1). The contribution of any other country into global volume of publications in UAVs was less than 5% for 1985 – 2015.

In the Europe the leaders on number of publications are Western European countries – UK with 935 publications on UAVs for 1985 – 2015 (4.5% of global volume of publications), Germany with 872 publications (4.2%) and to a less extent France with 724 publications (3.5%). Among Eastern European countries the leader is Poland with 227 publications for 1985 – 2015 (1.1 % of contribution into global volume of publications). It is followed by the Russian Federation (169 publications, 0.8%), Czech Republic (116 publications, 0.6%) and Ukraine (99 publications).

In South America the leader is Brazil with 286 publications (1.4% global volume of publications) on UAVs for 1985 – 2015. It is followed by Argentina with 37 Publications. In Africa the leaders are Algeria – 67 publications on UAVs (0.3% of global volume of publications) and South Africa – 57 publications. In Asia the other important players (in addition to China) in UAV research are South Korea with 794 publications for 1985 – 2015 (3.9% of global volume of publications) and Japan with 603 publications (2.9%).

*Figure 2. Key countries in UAV research in Scopus in 1985 – 2015*

Notes: All types of documents and sources indexed in Scopus are included in the analysis. The analysis was done in 2nd decade of January 2016.

World average level of citations of UAV publications for 1985 – 2015 is 3.29 citations per one publication. In terms of citations on UAV publications, the domination of the USA was much higher than in terms of publication activity. USA received 66.4% of all citations on UAV publications. The average level of citation of publications of the USA is 6.10 cites per document. China received 15.1% of world volume of publications, while UK 9.1%. Meanwhile in China the level of citation was much lower than in the USA – only 2.43 cites per document. In general, countries with the highest level of citations per one UAV publication have low level

\(^7\) Such a situation of the rocket growth of publication activity in China can be seen not only in UAVs studies but also in many other fields of science (see Kotsemir 2012a, 2012b).
of self-citation and quite high level of integration in international scientific collaboration (Figure 3).

Figure 3. Countries with the highest level of average citation in unmanned aerial vehicle research

Notes: All types of documents and sources indexed in Scopus are included in the analysis. The analysis was done in 2\textsuperscript{nd} decade of January 2016.

The highest level of citation per one publication ("cites per doc" further) has Switzerland: 11.24. In Hong Kong the value of this indicator was also very high: 10.25 cites per doc. Among countries with more than 50 publications on UAVs in Scopus for 1985 – 2015 the lowest "cites per doc" ratio has Ukraine – only 0.33 citations per publication. In some countries cites per doc ratio was lower than 1.00: Romania (0.79), Russian Federation (0.89) and Pakistan (0.94)\footnote{The problem of very low relative levels of citation in Russia, low level of citation in Asia vs. Europe, very high level of citation in Switzerland can also be seen for the total volume of publications both in Scopus and Web of Science databases (see Kotsemir 2012a, Kotsemir 2012b).}. In general, Western European countries have the highest level of citation per publication among macro-regions\footnote{This phenomenon can be seen for the total number of publications (see Kotsemir 2012a, Kotsemir 2012b).}. In the Asian region the highest level of cites per doc have countries with quite small number of UAV publications. It is already mentioned Hong Kong (with 10.25 "sites per doc" value; 61 publications) and to a lesser extent Israel (7.14; 170) and Singapore (6.58; 257). For comparison China, the second major player in UAV research with 4201 publications, has only 2.43 citations per one publication in average. Other Asian countries with big number of UAVs publications have low value of "cites per doc" indicator: Republic of Korea (3.78 with 794 publications); Japan (4.09 with 603 publications). Russia and Ukraine have the highest level of self-citation among countries with 50 and more publications on UAVs. 49.7\% of publications of Russia on UAVs was cited by authors of these publications. For Ukraine this figure is 44.7\%. Quite high share of self-citation has Finland (34.0\%) (see Table A.2 for details).
3. International collaboration in unmanned aerial vehicle research

Figure 4 depicts the map of cross-country collaborations in unmanned aerial vehicle research in Scopus for 2000 – 2015. This map was plotted using VOSviewer software. The dominance of the United States and China is clearly seen on this map. The United States can be treated as metaintegrator on this co-authorship map due to its dominating role in UAVs research.

Figure 4. Collaboration map of countries in the field of unmanned aerial vehicle studies in Scopus for 2000 – 2015, coloring based on geographical location

Notes: All types of documents and sources indexed in Scopus are included in the analysis. The analysis was done in 2nd decade of January 2016. Country statistics is taken through proceeding Countries are colored based on geographical location: Green bubbles – Asia; Blue – Oceania; Magenta (dark pink) – North America; Violet – Europe; Olive – Africa; Orange – South America

We also can see other key players on this map such as UK, Australia, Japan, South Korea, France, Germany, Italy and Brazil. There are two clusters of collaboration: Asian cluster with the dominance of China, South Korea and Japan European cluster, where key nodes are Italy, Spain, Germany, and France. Australia is included into intercontinental collaboration: cooperation with UK, France, USA, Canada, South Korea and China have high importance for Australia. It should be noted nevertheless that Asian partners for Australia are far less important for Australia than European and North American ones.

According to description from it official website: “VOSviewer is a software tool for constructing and visualizing bibliometric networks. These networks may for instance include journals, researchers, or individual publications, and they can be constructed based on co-citation, bibliographic coupling, or co-authorship relations”. VOSviewer also offers text mining functionality that can be used to construct and visualize co-occurrence networks of important terms extracted from a body of scientific literature”. Its names VOSviewer takes from (Visualization Of Similarities viewer). See more on http://www.vosviewer.com.
China with 3299 publications on unmanned aerial vehicles in Scopus in 2000-2015\textsuperscript{11} has only 354 co-authorship links. China takes here "publish or perish" strategy. Chinese authors are publishing very actively in the field of unmanned aerial vehicles primarily with each other without active (in comparison to the whole volume of publications on UAVs) integration into international scientific collaboration. Other leading Asian countries in UAV research also have low share of integration in international collaboration. India has 77 co-authorship links per 303 publications Republic of Korea has 174 co-authorship links per 646 publications. In leading European countries the level of integration into international scientific collaboration is much higher, than un Asia. United Kingdom has 353 co-authorship links per 786 publications, Germany – 317 co-authorship links per 742 publications, France – 408 co-authorship links per 629 publications.

4. Key organisation in UAV studies

The analysis of publication activity for organisations\textsuperscript{12} with the biggest number of UAVs publication is synthesized in Table A.3. US and Chinese organisations hit global ranking of centers of excellence on absolute number of UAV publications (Figure 5).

![Figure 5. Key centers of excellence on UAVs research in Scopus in 1985 – 2015.](image)

Notes: All types of documents and sources indexed in Scopus are included in the analysis. The analysis was done in 2\textsuperscript{nd} decade of January 2016.

The most productive organisation in UAVs studies in Scopus in 1985 – 2015 is Northwestern Polytechnical University (China, Xi'an) with 592 publications (14.1\% of contribution into total number of Chinese publications). It is followed by: Northwestern Polytechnical University (China, Xi'an) with 457 publications on UAVs (10.9\% of contribution into total number

\textsuperscript{11} Since VOSviewer detects country names from bibliographical descriptions of publications more accurately than Scopus automatic affiliation matching mechanisms the number of publications for each country after proceeding their descriptions trough VOSviewer software is higher than when user takes country data directly from Scopus database.

\textsuperscript{12} We should note here that due to inaccurate indexation of organisation affiliation name by Scopus search engine some publications of a given organization are not linked with this organization automatically. This problem is more serious in non-English speaking countries (see Kotsemir and Shashnov, 2015 on more details on this problem). In other words, Northwestern Polytechnical University (China, Xi'an) in fact has more than 457 publication on UAVs in Scopus for 1985 – 2015.

142
of Chinese publications); Wright-Patterson AFB (Air Forces Base) (USA, Dayton) with 442 publications (6.0% of contribution into total number of US publications); Nanjing University of Aeronautics and Astronautics (China, Nanjing) with 428 publications (10.2% of contribution into total number of US publications). The highest level of citations per one UAV publication among organisations with 75 and more UAV publications for 1985 – 2015 have UC (University of California) in Berkeley (USA, Berkeley): 23.2 cites per doc and University of Pennsylvania (USA, Philadelphia): 22.7 cites per doc. Chinese organisations dominate in the global ranking of organisations on number of publications on UAVs. Meanwhile, they have low level of citations (Figure 5). This figure also show that among the top countries on UAVs studies US organisations have in general higher level of citations per one publication. Table A.3 also helps to detect the key national centers of excellence in UAV studies. E.g. in Switzerland 62.0% of national volume of UAVs publications was done by Eidgenossische Technische Hochschule Zurich (163 publications on UAVs for 1985 – 2015). In Singapore the main national centres of excellence on UAV studies are National University of Singapore (54.1% of contribution into national volume of UAV publications) and National Cheng Kung University (34.2%) (Table A.3). Linkopings Universitet with 90 publications contributed 44.1% of the total number of UAV publications in Sweden.

5. General findings and conclusions

This paper proposed the bibliometric overview of unmanned aerial vehicle research in different countries. The analysis shows the strong domination of the USA in terms of publication activity and number of citations received. Meanwhile we can see the growing importance of China in the last years. By 2015 China became the second greatest country in UAV research. Meanwhile in terms of citations European countries as well as USA and Canada take the leadership while Asian countries (except of small countries like Singapore and Hong Kong) lags behind the leaders. On the level of organisations the US and Chinese organisation are the most productive in UAV research. Meanwhile leading Chinese organisations have much smaller number of citations per one publication than their “competitors” from the USA, Canada and Western Europe.

Several possible directions of further research can be proposed. One of the possible ways for development of this research is a more detailed analysis of international collaboration in UAV studies with focus on individual organisations. The other potential development of this research is running the semantic analysis. Using VOSviewer or the other similar software, we can derive different terms from titles, keywords and abstracts of publications on unmanned aerial vehicles. Further using the algorithms of natural language processing and cluster analysis based on co-occurrences of terms and words in titles, abstracts and keywords of UAV publications most important topics of UAV research can be detected.

Acknowledgement

The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of the subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program.

Author Information: Maxim Kotsemir, Junior Research Fellow, Quantitative Modelling Unit, Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics, 20 Myasnitskaya street, 101000, Moscow, Russian Federation E-mail: mkontsemeir@hse.ru (corresponding author); phone:+7(495) 772-9590*11740; ORCID: 0000-0002-8286-4480. Scopus AuthorID: 55903946200. Personal page: https://www.hse.ru/en/staff/Maxim_Kotsemir#sci Author page in Social Science Research network (SSRN): https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=1989392
References


## Appendix

Table A.1. Dynamics of publication activity on UAVs research in the 25 leading countries on number of publications on UAVs in Scopus in 1985 – 2015

<table>
<thead>
<tr>
<th>№</th>
<th>Country</th>
<th>Number of publications in Scopus</th>
<th>Share in the global volume of publications in Scopus, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>United States</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>China</td>
<td>2</td>
<td>93</td>
</tr>
<tr>
<td>3</td>
<td>United Kingdom</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>Germany</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>South Korea</td>
<td>23</td>
<td>62</td>
</tr>
<tr>
<td>6</td>
<td>Australia</td>
<td>5</td>
<td>26</td>
</tr>
<tr>
<td>7</td>
<td>France</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>Italy</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>Canada</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>Japan</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>Spain</td>
<td>3</td>
<td>35</td>
</tr>
<tr>
<td>12</td>
<td>India</td>
<td>10</td>
<td>29</td>
</tr>
<tr>
<td>13</td>
<td>Brazil</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>Taiwan</td>
<td>4</td>
<td>30</td>
</tr>
<tr>
<td>15</td>
<td>Switzerland</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>Netherlands</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>Singapore</td>
<td>36</td>
<td>22</td>
</tr>
<tr>
<td>18</td>
<td>Turkey</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>19</td>
<td>Malaysia</td>
<td>20</td>
<td>37</td>
</tr>
<tr>
<td>20</td>
<td>Poland</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>21</td>
<td>Portugal</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>22</td>
<td>Sweden</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>23</td>
<td>Mexico</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>Israel</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>Russian Federation</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

Notes: All types of documents and sources indexed in Scopus are included in the analysis. The analysis was done in 2nd decade of January 2016.
Table A.2. Key indicators of publication activity on UAVs in leading countries in 1985-2015

<table>
<thead>
<tr>
<th>Country</th>
<th>Absolute number of publications</th>
<th>Volume of publications,</th>
<th>Average number of publications (cites per doc)</th>
<th>Share of self-citation, %</th>
<th>Share of non-cited publications, %</th>
<th>Number of internationally collaborated publications (ICP)</th>
<th>Share of internationally collaborated publications (ICP), %</th>
<th>Citations per doc for ICP</th>
<th>Citations per doc for purely national publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>7379</td>
<td>35.8</td>
<td>6.10</td>
<td>13.6</td>
<td>53.5</td>
<td>1,050</td>
<td>14.2</td>
<td>9.44</td>
<td>1.70</td>
</tr>
<tr>
<td>China</td>
<td>4201</td>
<td>20.4</td>
<td>2.43</td>
<td>18.2</td>
<td>55.5</td>
<td>410</td>
<td>9.8</td>
<td>5.44</td>
<td>2.58</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>935</td>
<td>4.5</td>
<td>6.62</td>
<td>16.6</td>
<td>47.6</td>
<td>345</td>
<td>36.9</td>
<td>7.12</td>
<td>1.12</td>
</tr>
<tr>
<td>Germany</td>
<td>872</td>
<td>4.2</td>
<td>4.09</td>
<td>23.8</td>
<td>47.8</td>
<td>267</td>
<td>30.6</td>
<td>5.91</td>
<td>1.80</td>
</tr>
<tr>
<td>South Korea</td>
<td>794</td>
<td>3.9</td>
<td>3.78</td>
<td>21.7</td>
<td>50.3</td>
<td>180</td>
<td>22.7</td>
<td>6.36</td>
<td>2.11</td>
</tr>
<tr>
<td>Australia</td>
<td>771</td>
<td>3.7</td>
<td>6.42</td>
<td>14.5</td>
<td>45.3</td>
<td>253</td>
<td>32.8</td>
<td>9.98</td>
<td>2.13</td>
</tr>
<tr>
<td>France</td>
<td>724</td>
<td>3.5</td>
<td>7.08</td>
<td>16.2</td>
<td>47.8</td>
<td>352</td>
<td>48.6</td>
<td>8.13</td>
<td>1.33</td>
</tr>
<tr>
<td>Italy</td>
<td>657</td>
<td>3.2</td>
<td>5.58</td>
<td>24.9</td>
<td>44.7</td>
<td>213</td>
<td>32.4</td>
<td>7.64</td>
<td>1.66</td>
</tr>
<tr>
<td>Canada</td>
<td>632</td>
<td>3.1</td>
<td>5.35</td>
<td>17.1</td>
<td>46.5</td>
<td>210</td>
<td>33.2</td>
<td>7.21</td>
<td>1.63</td>
</tr>
<tr>
<td>Japan</td>
<td>603</td>
<td>2.9</td>
<td>4.09</td>
<td>17.6</td>
<td>56.6</td>
<td>136</td>
<td>22.6</td>
<td>8.87</td>
<td>3.29</td>
</tr>
<tr>
<td>Spain</td>
<td>507</td>
<td>2.5</td>
<td>7.20</td>
<td>21.8</td>
<td>42.0</td>
<td>171</td>
<td>33.7</td>
<td>5.97</td>
<td>0.76</td>
</tr>
<tr>
<td>India</td>
<td>360</td>
<td>1.7</td>
<td>2.33</td>
<td>15.7</td>
<td>66.9</td>
<td>79</td>
<td>21.9</td>
<td>4.66</td>
<td>2.78</td>
</tr>
<tr>
<td>Brazil</td>
<td>286</td>
<td>1.4</td>
<td>2.80</td>
<td>26.3</td>
<td>63.6</td>
<td>101</td>
<td>35.3</td>
<td>5.51</td>
<td>4.16</td>
</tr>
<tr>
<td>Taiwan</td>
<td>285</td>
<td>1.4</td>
<td>3.53</td>
<td>20.2</td>
<td>53.7</td>
<td>30</td>
<td>10.5</td>
<td>4.00</td>
<td>1.15</td>
</tr>
<tr>
<td>Switzerland</td>
<td>263</td>
<td>1.3</td>
<td>11.54</td>
<td>13.3</td>
<td>33.8</td>
<td>100</td>
<td>38.0</td>
<td>10.37</td>
<td>0.85</td>
</tr>
<tr>
<td>Netherlands</td>
<td>259</td>
<td>1.3</td>
<td>2.93</td>
<td>25.8</td>
<td>54.1</td>
<td>127</td>
<td>49.0</td>
<td>3.11</td>
<td>1.12</td>
</tr>
<tr>
<td>Singapore</td>
<td>257</td>
<td>1.2</td>
<td>6.58</td>
<td>16.1</td>
<td>40.9</td>
<td>81</td>
<td>31.5</td>
<td>7.44</td>
<td>1.21</td>
</tr>
<tr>
<td>Turkey</td>
<td>254</td>
<td>1.2</td>
<td>4.62</td>
<td>12.3</td>
<td>54.3</td>
<td>35</td>
<td>13.8</td>
<td>8.03</td>
<td>1.97</td>
</tr>
<tr>
<td>Malaysia</td>
<td>242</td>
<td>1.2</td>
<td>1.49</td>
<td>25.8</td>
<td>61.6</td>
<td>49</td>
<td>20.2</td>
<td>1.20</td>
<td>0.77</td>
</tr>
<tr>
<td>Poland</td>
<td>227</td>
<td>1.1</td>
<td>2.09</td>
<td>28.1</td>
<td>61.7</td>
<td>29</td>
<td>12.8</td>
<td>7.45</td>
<td>5.72</td>
</tr>
<tr>
<td>Portugal</td>
<td>213</td>
<td>1.0</td>
<td>5.38</td>
<td>23.5</td>
<td>50.2</td>
<td>123</td>
<td>57.7</td>
<td>8.32</td>
<td>3.78</td>
</tr>
<tr>
<td>Sweden</td>
<td>204</td>
<td>1.0</td>
<td>5.63</td>
<td>16.1</td>
<td>45.1</td>
<td>72</td>
<td>35.3</td>
<td>7.79</td>
<td>1.61</td>
</tr>
<tr>
<td>Mexico</td>
<td>184</td>
<td>0.9</td>
<td>3.89</td>
<td>14.8</td>
<td>53.8</td>
<td>102</td>
<td>55.4</td>
<td>5.49</td>
<td>1.93</td>
</tr>
<tr>
<td>Israel</td>
<td>170</td>
<td>0.8</td>
<td>7.14</td>
<td>20.5</td>
<td>52.9</td>
<td>58</td>
<td>34.1</td>
<td>6.22</td>
<td>0.73</td>
</tr>
<tr>
<td>Russian Federation</td>
<td>169</td>
<td>0.8</td>
<td>0.89</td>
<td>49.7</td>
<td>69.2</td>
<td>29</td>
<td>17.2</td>
<td>1.55</td>
<td>2.05</td>
</tr>
<tr>
<td>Iran</td>
<td>158</td>
<td>0.8</td>
<td>3.20</td>
<td>16.8</td>
<td>60.8</td>
<td>19</td>
<td>12.0</td>
<td>2.47</td>
<td>0.70</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>116</td>
<td>0.6</td>
<td>1.23</td>
<td>27.2</td>
<td>63.8</td>
<td>25</td>
<td>21.6</td>
<td>2.60</td>
<td>1.82</td>
</tr>
<tr>
<td>Ukraine</td>
<td>99</td>
<td>0.5</td>
<td>0.33</td>
<td>44.7</td>
<td>79.8</td>
<td>6</td>
<td>6.1</td>
<td>0.33</td>
<td>0.86</td>
</tr>
<tr>
<td>Norway</td>
<td>99</td>
<td>0.5</td>
<td>2.36</td>
<td>30.8</td>
<td>57.6</td>
<td>46</td>
<td>46.5</td>
<td>3.09</td>
<td>1.78</td>
</tr>
<tr>
<td>Greece</td>
<td>98</td>
<td>0.5</td>
<td>6.86</td>
<td>16.5</td>
<td>50.0</td>
<td>48</td>
<td>49.0</td>
<td>9.88</td>
<td>2.41</td>
</tr>
<tr>
<td>Austria</td>
<td>95</td>
<td>0.5</td>
<td>3.48</td>
<td>16.7</td>
<td>38.9</td>
<td>36</td>
<td>37.9</td>
<td>5.00</td>
<td>1.83</td>
</tr>
<tr>
<td>Belgium</td>
<td>93</td>
<td>0.5</td>
<td>6.83</td>
<td>19.3</td>
<td>54.8</td>
<td>47</td>
<td>50.5</td>
<td>10.79</td>
<td>3.49</td>
</tr>
<tr>
<td>Hungary</td>
<td>87</td>
<td>0.4</td>
<td>2.91</td>
<td>31.4</td>
<td>59.8</td>
<td>14</td>
<td>16.1</td>
<td>13.21</td>
<td>11.22</td>
</tr>
<tr>
<td>Pakistan</td>
<td>73</td>
<td>0.4</td>
<td>0.94</td>
<td>12.2</td>
<td>61.6</td>
<td>17</td>
<td>23.3</td>
<td>0.53</td>
<td>0.41</td>
</tr>
<tr>
<td>Romania</td>
<td>73</td>
<td>0.4</td>
<td>0.79</td>
<td>25.9</td>
<td>71.2</td>
<td>15</td>
<td>20.5</td>
<td>1.53</td>
<td>2.54</td>
</tr>
<tr>
<td>United Arab Emirates</td>
<td>73</td>
<td>0.4</td>
<td>1.85</td>
<td>20.0</td>
<td>54.8</td>
<td>46</td>
<td>63.0</td>
<td>2.39</td>
<td>2.58</td>
</tr>
<tr>
<td>Algeria</td>
<td>67</td>
<td>0.3</td>
<td>6.67</td>
<td>4.3</td>
<td>49.3</td>
<td>32</td>
<td>47.8</td>
<td>13.00</td>
<td>6.41</td>
</tr>
<tr>
<td>Finland</td>
<td>63</td>
<td>0.3</td>
<td>5.57</td>
<td>34.0</td>
<td>46.0</td>
<td>22</td>
<td>34.9</td>
<td>2.05</td>
<td>0.26</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>61</td>
<td>0.3</td>
<td>10.25</td>
<td>13.8</td>
<td>36.1</td>
<td>50</td>
<td>82.0</td>
<td>11.32</td>
<td>1.81</td>
</tr>
<tr>
<td>South Africa</td>
<td>57</td>
<td>0.3</td>
<td>2.87</td>
<td>12.0</td>
<td>66.7</td>
<td>12</td>
<td>21.1</td>
<td>8.75</td>
<td>5.63</td>
</tr>
</tbody>
</table>

Notes: All types of documents and sources indexed in Scopus are included in the analysis. The analysis was done in 2nd decade of January 2016.

<table>
<thead>
<tr>
<th>Affiliation name (as it named in Scopus)</th>
<th>Total number of documents (publication) 1985 – 2015</th>
<th>Total number of citations 1985 – 2015</th>
<th>Average number of citations per one document (“cites per doc”)</th>
<th>Share in national number of publications, %</th>
<th>Organisation “cites per doc”/National “cites per doc” ratio</th>
<th>Share of self citations, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beihang University, China, Beijing</td>
<td>592</td>
<td>1343</td>
<td>2.27</td>
<td>14.1</td>
<td>0.93</td>
<td>22.8</td>
</tr>
<tr>
<td>Northwestern Polytechnical University, China, Xi’an</td>
<td>457</td>
<td>587</td>
<td>1.28</td>
<td>10.9</td>
<td>0.53</td>
<td>18.2</td>
</tr>
<tr>
<td>Wright-Patterson AFB, USA, Dayton</td>
<td>442</td>
<td>2155</td>
<td>4.88</td>
<td>6.0</td>
<td>0.80</td>
<td>17.2</td>
</tr>
<tr>
<td>Nanjing University of Aeronautics and Astronautics, China, Nanjing</td>
<td>428</td>
<td>1033</td>
<td>2.41</td>
<td>10.2</td>
<td>0.99</td>
<td>20.2</td>
</tr>
<tr>
<td>Georgia Institute of Technology, USA, Atlanta</td>
<td>243</td>
<td>1105</td>
<td>4.55</td>
<td>3.3</td>
<td>0.75</td>
<td>23.8</td>
</tr>
<tr>
<td>University of Florida, USA, Gainesville</td>
<td>234</td>
<td>2428</td>
<td>10.38</td>
<td>3.2</td>
<td>1.70</td>
<td>20.1</td>
</tr>
<tr>
<td>Massachusetts Institute of Technology, USA, Cambridge</td>
<td>217</td>
<td>1973</td>
<td>9.09</td>
<td>2.9</td>
<td>1.49</td>
<td>16.0</td>
</tr>
<tr>
<td>University of Maryland, USA, College Park</td>
<td>213</td>
<td>1568</td>
<td>7.36</td>
<td>2.9</td>
<td>1.21</td>
<td>24.3</td>
</tr>
<tr>
<td>Cranfield University**, UK, Cranfield</td>
<td>213</td>
<td>1184</td>
<td>5.56</td>
<td>22.8</td>
<td>0.84</td>
<td>19.2</td>
</tr>
<tr>
<td>Tsinghua University, China, Beijing</td>
<td>189</td>
<td>659</td>
<td>4.54</td>
<td>4.5</td>
<td>1.87</td>
<td>14.6</td>
</tr>
<tr>
<td>Beijing Institute of Technology, China, Beijing</td>
<td>186</td>
<td>228</td>
<td>1.23</td>
<td>4.4</td>
<td>0.50</td>
<td>16.7</td>
</tr>
<tr>
<td>National University of Defense Technology, China, Changsha</td>
<td>181</td>
<td>293</td>
<td>1.62</td>
<td>4.3</td>
<td>0.67</td>
<td>11.3</td>
</tr>
<tr>
<td>Brigham Young University, USA, Provo</td>
<td>177</td>
<td>2794</td>
<td>15.79</td>
<td>2.4</td>
<td>2.59</td>
<td>10.8</td>
</tr>
<tr>
<td>Virginia Polytechnic Institute and State University, USA, Blacksburg</td>
<td>168</td>
<td>1179</td>
<td>7.02</td>
<td>2.3</td>
<td>1.15</td>
<td>26.1</td>
</tr>
<tr>
<td>Eidgenossische Technische Hochschule Zürich**, Switzerland, Zurich</td>
<td>163</td>
<td>1730</td>
<td>10.61</td>
<td>62.0</td>
<td>0.92</td>
<td>16.1</td>
</tr>
<tr>
<td>Korea Advanced Institute of Science &amp; Technology, South Korea, Daejon</td>
<td>161</td>
<td>732</td>
<td>4.55</td>
<td>20.3</td>
<td>1.20</td>
<td>18.2</td>
</tr>
<tr>
<td>Seoul National University, South Korea, Seoul</td>
<td>144</td>
<td>696</td>
<td>4.83</td>
<td>18.1</td>
<td>1.28</td>
<td>13.5</td>
</tr>
<tr>
<td>UC Berkeley, USA, Berkeley</td>
<td>143</td>
<td>3315</td>
<td>23.18</td>
<td>1.9</td>
<td>3.80</td>
<td>9.4</td>
</tr>
<tr>
<td>National University of Singapore**, Singapore, Singapore City</td>
<td>139</td>
<td>998</td>
<td>7.18</td>
<td>54.1</td>
<td>1.09</td>
<td>16.2</td>
</tr>
<tr>
<td>University of Colorado at Boulder, usa, Boulder</td>
<td>136</td>
<td>726</td>
<td>5.34</td>
<td>1.8</td>
<td>0.88</td>
<td>20.5</td>
</tr>
<tr>
<td>University Michigan Ann Arbor, usa, Ann Arbor</td>
<td>134</td>
<td>1424</td>
<td>10.63</td>
<td>1.8</td>
<td>1.74</td>
<td>19.8</td>
</tr>
<tr>
<td>The University of Sydney, Australia, Sydney</td>
<td>132</td>
<td>838</td>
<td>6.35</td>
<td>17.1</td>
<td>0.99</td>
<td>12.6</td>
</tr>
<tr>
<td>Delft University of Technology**, Netherlands, Delft</td>
<td>130</td>
<td>355</td>
<td>2.73</td>
<td>50.2</td>
<td>0.93</td>
<td>24.5</td>
</tr>
<tr>
<td>NASA Ames Research Center, USA, Moffett Field</td>
<td>127</td>
<td>548</td>
<td>4.31</td>
<td>1.7</td>
<td>0.71</td>
<td>21.0</td>
</tr>
<tr>
<td>Naval Postgraduate School, USA, Monterey</td>
<td>121</td>
<td>921</td>
<td>7.61</td>
<td>1.6</td>
<td>1.25</td>
<td>17.9</td>
</tr>
<tr>
<td>Chinese Academy of Sciences, China, Beijing</td>
<td>112</td>
<td>367</td>
<td>3.28</td>
<td>2.7</td>
<td>1.35</td>
<td>27.0</td>
</tr>
<tr>
<td>Boeing Corporation, USA, Chicago</td>
<td>112</td>
<td>525</td>
<td>4.69</td>
<td>1.5</td>
<td>0.77</td>
<td>18.1</td>
</tr>
<tr>
<td>Naval Research Laboratory, USA, Washington</td>
<td>110</td>
<td>1036</td>
<td>9.42</td>
<td>1.5</td>
<td>1.54</td>
<td>18.2</td>
</tr>
<tr>
<td>Air Force Engineering University China, China, Xi’an</td>
<td>110</td>
<td>157</td>
<td>1.43</td>
<td>2.6</td>
<td>0.59</td>
<td>21.7</td>
</tr>
<tr>
<td>Deutsches Zentrum fur Luft- Und Raumfahrt, Germany, Cologne</td>
<td>107</td>
<td>335</td>
<td>3.13</td>
<td>12.3</td>
<td>0.77</td>
<td>23.0</td>
</tr>
<tr>
<td>Universite Concordia, Canada, Montreal</td>
<td>106</td>
<td>352</td>
<td>3.32</td>
<td>16.8</td>
<td>0.62</td>
<td>25.3</td>
</tr>
<tr>
<td>Carnegie Mellon University, USA, Pittsburgh</td>
<td>105</td>
<td>896</td>
<td>8.53</td>
<td>1.4</td>
<td>1.40</td>
<td>10.9</td>
</tr>
<tr>
<td>Royal Melbourne Institute of Technology University, Australia, Melbourne</td>
<td>103</td>
<td>380</td>
<td>3.69</td>
<td>13.4</td>
<td>0.57</td>
<td>20.8</td>
</tr>
<tr>
<td>Korea Aerospace Research Institute, South Korea, Daejon</td>
<td>101</td>
<td>240</td>
<td>2.38</td>
<td>12.7</td>
<td>0.63</td>
<td>17.9</td>
</tr>
</tbody>
</table>

Notes: All types of documents and sources indexed in Scopus are included in the analysis. The analysis was done in 2nd decade of January 2016. In bold key centers of excellence of a country are highlighted.
A parameter-free index for identifying under-cited sleeping beauties in science

Du Jian¹ Wu Yishan²

¹windowsdujian@163.com
Institute of Medical information & Library, Chinese Academy of Medical Sciences, Beijing(China)

²wuyishan1958@163.com
Chinese Academy of Science and Technology for Development, Beijing(China)

Abstract

We developed a systematic methodology for identifying the under-cited (or not-so-highly cited) Sleeping Beauty (SB) publications and tried to figure out their key characteristics. Based on the identification framework of “beauty coefficient” (B) introduced by Ke et al. (2015), taking into account the whole citation history of the publications concerned, we substituted yearly citations in “beauty coefficient” with yearly accumulative percentage of citations, and eliminated the denominator in “beauty coefficient” since the curve of a given document’s accumulative citations is always monotonically increasing if only the document is cited. The value of the modified beauty coefficient is denoted as Bcp. We also redefined the awakening year, sleeping length and sleeping depth within the Bcp framework with the intention of avoiding arbitrary thresholds as much as possible. We tested the new index using the data of SB articles identified from Science and Nature. The results showed that Bcp is more sensitive in identifying the “lower level SBs”, which refers to the case when the total citations and the maximum annual citations of SBs are not so high in comparison with other typical SBs. Bcp works better than B in at least two aspects: (1) it “punishes” the situations when the SBs experienced early citations instead of continuous sleeping; (2) it allows for comparing the extent of delayed citation impact of publications in different fields with different citation patterns. We also figured out some key characteristics of such SB publications and pondered some policy implications about the associations of SB publications with transformative research, research front and research evaluation.

Conference Topic

Indicators
Science policy and research assessment
Citation and co-citation analysis

Introduction

A “Sleeping Beauty” (SB) in Science is a publication that goes unnoticed (or “sleeps”) for a long time and then, almost suddenly, attracts a lot of attention, or “is awakened by a Prince” (van Raan, 2004). This concept is actually a quantitative description of “delayed recognition of scientific achievements”, a phenomenon widely discussed in sociology of science (Garfield, 1989). “Premature discoveries” and “transformative innovations” are crucial for the development of science, but they are often initially neglected or resisted by the scientific community and thus are often subject to delayed recognition (Trapido, 2015). In this paper, we try to propose a systematic identification method of SB publications, in order to extend the application scope of citation analysis, and discussed implications for identifying potential “ahead of time” discoveries or transformative research, and shorten time lag for original research to get recognized.

There are three types of methods for identifying SBs in science, i.e., citation curve fitting, arbitrary thresholds setting, and parameter-free index, which were described at full length in our foregoing paper (Du & Wu, 2016). Each of the three methods has both advantages and disadvantages.
Firstly, it is not precise and accurate for identifying SB publications with citation curve fitting since it is inefficient to mine massive publications based on manual observation.

Secondly, the arbitrary thresholds methods include average-based and quartile-based criteria, as well as their combinations. Average-based criteria directly or indirectly referring to van Raan’s definitions on sleeping period, sleeping depth, awakening period and awakening intensity (van Raan, 2004), could better define the never cited or poorly cited papers in sleeping period. While quartile-based criteria could better reflect the extent of citation delay in the entire citation life time of an article. The former is based on a more strict threshold, and the latter a more moderate threshold. But it is statistically difficult to operate for the latter since one paper’s citation history should be compared with that of all papers across the research fields (Costas, Van Leeuwen, & Van Raan, 2010). Both the average-based and quartile-based criteria are arbitrary definitions and do not take into account different citation patterns in various research fields. For example, van Raan defined four main variables, (1) length of the sleep in years after publication; (2) depth of sleep in terms of a maximum citation rate during the sleeping period; (3) awake period in years after the sleeping period; and (4) awake intensity in terms of a minimum citation rate during the awake period. The four main variables can be tuned. He denoted the SBs with these variables with [5,1,4,5] in (van Raan, 2004) and [10,1,10,5] in (van Raan, 2015, 2017). The combination of average-based with quartile-based criteria enhanced the accuracy for identifying sleeping beauties, but the combined method increased complexity and reduced transparency.

Finally, the two main parameter-free index include Citation Delay (Wang, Thijs, & Glanzel, 2015), a contrary indicator of Citation Speed (Wang, 2013), and Beauty Coefficient (Ke, Ferrara, Radicchi, & Flammini, 2015). In our foregoing paper (Du & Wu, 2015), these two parameter-free indices were used to identify SBs published between 1970 and 2005 in the four most prestigious clinical medicine journals, i.e., New England Journal of Medicine, The Lancet, Journal of the American Medical Association and British Medical Journal. We found that because Citation Speed takes the citation curve of the whole citation window into account, by it one could identify the continuingly highly cited papers with "very long" life cycle but could not directly distinguish SB publications. Beauty Coefficient was proved to be a good indicator to identify a SB, but it fails to cover the citation curve after the paper receives its maximum annual citations. Nevertheless, considering three factors including scientifically designed, transparently calculated and simply applied, the Beauty Coefficient is still a relatively better measure.

(Li & Ye, 2016) proposed four rules that should be adhered to in distinguishing SBs in science: (1) the early citations should be penalized; (2) the whole citation history should be taken into account; (3) the awakening time of a sleeping beauty should not vary over time; and (4) arbitrary thresholds on sleeping period or awakening intensity should be avoided. In this paper, we tried to combine the framework of Citation Delay and Beauty Coefficient to propose a new parameter-free index for distinguishing the under-cited sleeping beauties in science.

**Methods and Materials**

**Beauty Coefficient, B**

Beauty Coefficient is based on the comparison between its citation history and a reference line, drawn from its publication year to the year of the peak of citations. Let's call \( t \) the time...
interval after publication and $c_t$ the citation history of the paper. If $c_{tm}$ is the maximum of $c_t$, the straight line $\ell_t$ that connects the point $(0, c_0)$ and $(t_m, c_{tm})$ analytically is described by the equation:

$$\ell_t = \frac{c_{tm} - c_0}{t_m} \cdot t + c_0$$

$(c_{tm} - c_0)/t_m$ is the slope of the line $\ell_t$. For each $t < t_m$ we can compute the ratio between $\ell_t - c_t$ and max{1, $c_t$}, and the definition of B is achieved by summing over these values.

$$B = \sum_{t=0}^{t_m} \frac{t_m}{t_m} \frac{c_{tm} - c_0}{t_m} \cdot t + c_0 - c_t \max\{1, c_t\}$$

The advantage of B is that it does not rely on arbitrary thresholds or on certain percentage. By it we can investigate this phenomenon at a systematic level. But we do not agree with (Ke et al., 2015)’s argument that B can be calculated for any given paper that received at least one citation. For example, for a given paper, if the maximum number of citations received just in the publication year and the yearly citations decreased in the following years, i.e., $c_0 = c_{tm}$, then the reference line does not exist, and thus the value B could not be calculated.

One of the disadvantages of this definition is the high importance given to the peak. B works really well with top class SBs that after discovery have huge numbers of citations every year, but for lower level SBs with less total citations, it gives some unwanted results: many of the papers we had found as SBs had very low B values.

The other obvious disadvantage of B is that the denominator max{1, $c_t$} does not penalize; instead, it may advance early citation accumulation. This runs against the intention of the authors (Ke et al., 2015). We hold that the role of the denominator is just to avoid division by zero, and calculating the ratio with a denominator could lead to the lost of the original information of the citation history.

### Modifying Beauty Coefficient

So, to avoid the dependence of Beauty Coefficient on just one year’s citation history and to reduce the sensitivity of the measure to such extreme SBs, we propose a new measure in order to discover the SBs in the not-so-highly cited publications. Such under cited publications were often ignored in bibliometric related research.

Based on the identification framework of “beauty coefficient” (B) introduced by Ke et al. (2015), taking into account the whole citation history of the publications concerned, we substituted yearly citations in “beauty coefficient” with yearly accumulative percentage of citations, and eliminated the denominator in “beauty coefficient” since the curve of a given document’s accumulative citations is always monotonically increasing if only the document is cited. The value of the modified beauty coefficient is denoted as Bcp. The new beauty coefficient value Bcp for a given paper is based on the comparison between its citation history of annual accumulative percentage and a reference line that is determined only by its publication year and the share of citations in this year, the maximum accumulative percentage of citations received in a year (100% within a multiyear observation period), and the year when such maximum is achieved:

$$Bcp = \sum_{t=0}^{t_m} \frac{1 - c_0}{t_m} \cdot t + c_0 - c_t$$
Similar with (Ke et al., 2015)’s definition, we gave a plausible definition of awakening time—the year when the abrupt change in the accumulation of citations of SBs occurs. We defined the awakening time $t_a$ as the time $t$ at which the distance $d(t)$ between the point $(t, c_t)$ and the reference line $t = t_m$ reaches its maximum:

$$d(t) = \frac{1 - c_0 \cdot t - c_t \cdot (1 - c_0)}{\sqrt{(1 - c_0)^2 + t_m^2}}$$

The awakening time reflect the year when the abrupt change in the accumulation of citations of SBs occurs, not the abrupt change from one year’s citations to other year’s citations. Being able to pinpoint the awakening time may help identify possible general trigger mechanisms behind the said change.

In Figure 1, the reference line denotes that a given paper’s annual citations is equal. Bcp > 0 denotes delayed citations and Bcp < 0 early citations. For a given paper, if the total number of citations received only in the publication year and the yearly citations after publication is zero, then the value of Bcp is -(n-1)/2, n is the age of a given paper. In contrast, if the total number of citations received only in the last year and the yearly citations after publication till the last year is zero, then the value of Bcp is (n-1)/2.

**Figure 1. Illustration of the definition of the new beauty coefficient and the awakening time of a paper.**

In should be noted that the framework of Bcp allows comparing the extent of delayed citation impact of publications in different disciplines with different citation patterns. Bcp depends on the relative shape of the graph shown in Fig. 1 but not on the total number of citations.

**Dataset**

Two parameter-free indices, i.e. Beauty Coefficient (B) and the modified Beauty Coefficient (Bcp) were used to identify sleeping beauties articles published between 1970 and 2005 in *Science* and *Nature.* The total period in which the SBs and their citation data are searched for
is 1970-2015. Thus, 2005 is the last year for publications having in total a ten year time span until 2015. Articles with at least 200 citations, in total 20,000 publications were included in the following analysis.

**Results**

*Bcp works better than B*

*A case study in (Li & Ye, 2016)*

As for the example shown in (Li & Ye, 2016), two papers P1 and P2 were published in the same year, received the same number of citations, but had different citation curves as shown in Fig. 2. Both of them received no citation in the publication year, and reached citation peak at the age of twenty. Citations of P2 were received earlier than those of P1.

![Figure 2. Assumed articles P1 and P2. The upper graph A and bottom graph B are under B framework (yearly citations) and Bcp framework (yearly cumulative percentage of citations), respectively](image)

In terms of beauty coefficient, they have the same reference line. It is calculated that $B_1 = 164.75$ is smaller than $B_2 = 177.95$, although P1 accumulated citations later than P2. It is strange! But within our Bcp framework, it is calculated that $Bcp_1 = 5.075$ is larger than $Bcp_2 = 1.075$. Obviously, P1 has a delayed citation impact than P2. From this case, we can see the new beauty coefficient is better than B. Next, we will validate the new index through a large-scale bibliographic dataset.
We selected top 1% (N=200) sleeping beauty publications with B and Bcp, respectively. There are 133 papers occurred in both top 1% list either by B or by Bcp, and 134 ((200-133)*2) papers only occurred in top 1% list by one index but not by the other index. So, we mainly analyze the 134 discrepant papers. Within our Bcp framework, we redefined two variables: (1) length of the sleep: years in the sleeping period, i.e., the time interval between publication year and awakening year, (2) depth of sleep: the accumulated percentage of citations in the year before the awakening year, that is the accumulated percentage of citations at the end of sleeping period. The difference test between B and Bcp is shown in Table 1.

<table>
<thead>
<tr>
<th>Index</th>
<th>N</th>
<th>Median</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>length of the sleep</td>
<td>B</td>
<td>67</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Bcp</td>
<td>67</td>
<td>20.1</td>
</tr>
<tr>
<td>depth of sleep</td>
<td>B</td>
<td>67</td>
<td>42.9</td>
</tr>
<tr>
<td></td>
<td>Bcp</td>
<td>67</td>
<td>28</td>
</tr>
<tr>
<td>total citations</td>
<td>B</td>
<td>67</td>
<td>3908</td>
</tr>
<tr>
<td></td>
<td>Bcp</td>
<td>67</td>
<td>1561.9</td>
</tr>
<tr>
<td>annual maximum citations</td>
<td>B</td>
<td>67</td>
<td>183.4</td>
</tr>
<tr>
<td></td>
<td>Bcp</td>
<td>67</td>
<td>112.6</td>
</tr>
</tbody>
</table>

Significant at the 0.05 level (2-tailed); ** Significant at the 0.01 level (2-tailed)

The average length of sleep of sleeping beauty publications identified by Bcp is significantly smaller than B. In other words, the awaking year identified with B framework is often delayed estimated than Bcp because of the computing mechanism. The average depth of sleep of sleeping beauty publications identified by Bcp is significantly smaller than B. That is to say, Bcp works better than B in penalizing the early citations. Although not statically significant, the total citations and annual maximum citations of sleeping beauty publications identified by Bcp is also smaller than B. Bcp is more sensitive for in identifying “the lower level SBs”, which refers to the case when the total citations and the maximum annual citations of SBs are not so high in comparison with other typical SBs.

We then selected top 0.1% (N=20) sleeping beauty publications with B and Bcp, respectively. There are 12 papers occurred in both top 1% list either by B or by Bcp, and 8 papers only occurred in top 1% list by one index but not by the other index. So, we mainly analyze the citation curve of the 8 discrepant papers. As is shown in Fig. 3, we found that in contrast with B, Bcp works well in penalizing early citations, so that, at parity of the total citations received, the later such citations accumulate, the higher the value of the proposed measure. Bcp takes into account the entire, rather than the partial, citation history of an article, so the extreme sleeping beauty publications identified with Bcp showed a continually increasing trend in terms of annual citations. Such articles demonstrated a delayed but durable citation impact.
The extreme SBs identified by Bcp tend to be the landmark publications of a specific research field and the SBs are often technique and application-oriented work.

As is shown in Table 2, in terms of content analysis, we could see that the top 10 SBs identified by Bcp were all the landmark publications of a specific research field, such as “the first report on …”, or “the classic theory about …”. Three papers were Nobel laureate’s publications. For example, John Maynard Smith’s concept of protein space proposed were assessed by Nature that “1970 Foreshadows concepts now widely applied in studies of molecular evolution, such as genotype-phenotype mapping” (http://www.nature.com/nature/focus/maynardsmith/). It appears that high quality publications tend to encounter delayed recognition and thus showed delayed citation impact.

One is perhaps more inclined to believe that Sleeping Beauties relate to more fundamental and basic, and less to application-oriented work. But a surprising finding is that half of the SBs are application oriented and significantly more cited in patents than ‘normal’ papers (van Raan, 2015, 2017). The scientific non-patent references (SNPRs) represent a bridge between science and technology although they do not necessarily indicate the direct scientific basis of the invention described in the patent. In this study we focus on a particular phenomenon, namely the extent to which the extreme SBs show up as SNPRs. Patent publications were gathered by searching the lens.org, created by Cambia (a non-profit organization in Australia dedicated to facilitating innovation) and Queensland University of Technology. The platform lens.org has linked the world's patent information to most of the scholarly literature with collaborations with CrossRef and National Library of Medicine. We group patent publications describing the same invention in ‘patent families’ to prevent double counting. As is shown in Table 2, six of the top 10 SBs identified by Bcp have been cited by patents, and the SB’s first citation in a patent usually appears to be earlier than the awakening year.
<table>
<thead>
<tr>
<th>Bcp</th>
<th>Title</th>
<th>Authors</th>
<th>Source</th>
<th>Fields</th>
<th>Landmark publications</th>
<th>Citations</th>
<th>Awakening year</th>
<th>Citations by patent families</th>
<th>Application date of the earliest citing patent</th>
<th>Priority date</th>
<th>Patent Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.360</td>
<td>Selection and Covariance</td>
<td>Price, GR</td>
<td>Nature, 1970, 227(5257), P520</td>
<td>Biology, Life Science</td>
<td>In the theory of evolution and natural selection, the Price equation (also known as Price's equation or Price's theorem) describes how a trait or gene changes in frequency over time.</td>
<td>736</td>
<td>2003</td>
<td>-</td>
<td>Sep 27, 2000</td>
<td>Sep 27, 2000</td>
<td>US 23604500</td>
</tr>
</tbody>
</table>

Note: Citations are counted through Web of Science between the publication year and 2015. Citations by patent families are searched for through lens.org from the publication year till 14 Feb, 2017.
Concluding remarks

In this paper, we developed a systematic methodology for identifying SB publications and figured out their key characteristics. We tested the new index using the data of SB articles identified from *Science* and *Nature*. The results showed that both B and Bcp worked well with top class SBs, but Bcp is more sensitive for the lower level SBs, i.e., the total citations and the maximum number of citations received in a year is moderate. Bcp works better than B in at least two aspects: (1) it “punishes” the situations when the SBs experienced early citations instead of continuous sleeping; (2) it allows for comparing the extent of delayed citation impact of publications in different fields with different citation patterns. We also figured out some key characteristics of SB publications. The extreme SBs are application oriented and significantly more cited in patents than ‘normal’ papers. The SB’s first citation in a patent usually appears to be earlier than the awakening year. These findings demonstrated the potential technical and application-research properties of Sleeping Beauties. Based on the above mentioned analysis, we pondered some policy implications associated with SB publications.

The first is SB publications and transformative research. According to Thomas Kuhn’s “paradigm” concept, there should be two types of innovative research, i.e., cumulative processes and revolutionary breakthroughs. The latter is also called transformative research, which refers to research that shifts or disrupts established scientific paradigms. Identifying potential transformative research early and accurately is important for funding agencies to maximize the impact of their investments. It also helps scientists identify promising emerging works and focus their attention on them. Therefore, it is imperative for us to spare no effort to avoid delayed recognition and to detect SBs as early as possible, in order to promote potentially valuable but not readily accepted innovative research. Transformative research tend to be neglected or resisted by the scientific community initially and this neglect or resistance could be regarded as the key clue for the early prediction of sleeping beauty literature. Publications belonging to so-called transformative research, even when less frequently cited than others, should be given special attention as early as possible, because they may suddenly attract many citations after a period of sleep. We hold that scholars in both scientometrics and library and information science (LIS) should initiate the research for identifying transformative ideas. One could identify transformative research through some text terms (such as "disagree", "overcome", "break", "dispute"...). In order to discern such potential transformative research, we could observe whether the relevant documents get early citation from patents or not, because many Sleeping Beauty documents tend to be technical research in nature and application-oriented.

The second concerns SB publications and research fronts. The research front(s) studies based on citation analysis and visualization methods were one of most important topics in informetrics. However, research front and research frontier are different terms, which translated into same Chinese Word. The former usually took the "fast-highly cited papers" as basic data, and latter tend to be revealed by the delay-highly cited papers from the perspective of scientists, as is shown in terms of the landmark characteristic of the top ten sleeping beauty publications identified with Bcp. The Bcp index proposed in this paper may be used to identify the papers at the sleeping-awakening interface, which provided new tools for arousing attention of the science community to previously overlooked but important research.

The third is SB publications and research evaluation. Since the major achievements often encounter delayed recognition than the hot tracking research, it is recommended that citation
delay reflected by Bcp can be used as an important index to evaluate the academic quality of papers and it is suggested to moderately extend the evaluation cycle. In bibliometric-based research assessment, one should give special attention to papers with a higher value of Bcp, because such papers accumulate their citations slowly and show a longer and durable citation impact.

Acknowledgments
This study was supported by the National Natural Science Foundation of China (Grant No. 71603280, 71373252).

References


How is CiteSpace used and cited in the literature? An analysis of the articles published in English and Chinese core journals

Xuelian Pan¹  Ming Cui ²  Xiaotong Yu ³  Weina Hua ⁴

¹_xuelianpan@nju.edu.cn
Nanjing University, Nanjing (China)

²_969010781@qq.com
Nanjing University, Nanjing (China)

³_981925271@qq.com
Nanjing University, Nanjing (China)

⁴_huawn@nju.edu.cn
Nanjing University, Nanjing (China)

Abstract
This study investigates the use and citation of CiteSpace, a freely available tool for analysing, detecting and visualizing trends and patterns in scientific literature, by examining how it is used and cited among the articles published in English and Chinese core journals. Results show that CiteSpace is widely used in China along with a substantial uncitedness. The number of Chinese articles using CiteSpace is obviously increasing, while the citation rate of CiteSpace is not increasing over time. Many Chinese authors do not provide sufficient information for identifying CiteSpace. Findings also show that CiteSpace used in English core journal articles is more likely to receive citations than that used in Chinese core journal articles. Moreover, our results demonstrate that there are significant differences in citation counts between sections containing CiteSpace.

Conference Topic
Citation and co-citation analysis
Science policy and research assessment

Introduction
Software is important to scientific research: it assists scientists to identify research questions, analyse data, visualize the results and disseminate knowledge; indeed, “just about every step of scientific work is affected by software” (Howison et al., 2015, p. 454). However, the academic value of software has long been undervalued and, even worse, has been ignored in the current publication-driven scientific reward system. Recent years have witnessed a tremendous growth in software which is freely available for academic use (Hannay et al., 2009; Huang et al., 2013). As the value of data is increasingly recognized and a considerable amount of freely available software packages are used in the scientific community (Howison & Bullard, 2016; Thelwall & Kousha, 2016), some scholars argue that software should also be valued as an academic contribution (Hafer & Kirkpatrick, 2009; Piwowar, 2013). The US National Science Foundation (NSF) has recognized software as a valid research output since 2013 (NSF, 2013). Software has also been listed as a scholarly contribution in the UK Research Excellence Framework 2014 (Research Excellence Framework, 2013). Yet, many funding institutions, policy makers and administrators have not recognized software as a valid type of research products (Piwowar, 2013). Therefore, measuring the impact of software is imperative, because which will enable us to have a better understanding of the value of software and help to incorporate software as an integral component in research evaluations and scholarly communication.
Although citation counts is widely used to assess the impact of journal articles and monographs (Kousha, Thelwall, & Rezaie, 2011; Song & Kim, 2013; Cartes-Velásquez & Manterola Delgado, 2014), it still needs to investigate the current status of software citation practice before using citation counts to measure the impact of software. A few studies examining the citation of software in scientific articles have found that software citation practice is not yet customary. Howison’s study on 90 biology articles has found that considerable software packages mentioned in the articles were not formally cited (Howison & Bullard, 2016). Our previous study on articles published in PLOS ONE has also found a similar result (Pan, Yan, & Hua, 2016). Therefore, it is worth investigating the usage of software in full-text scientific articles to demonstrate the academic impact software.

In this article, we examined how software was used and cited in different journal articles and factors affecting citation rate of software. CiteSpace (http://cluster.cis.drexel.edu/~cchen/citespace/; Chen, 2004, 2006), a freely available software tool for analysing, detecting and visualizing trends and patterns in scientific literature, was chosen as the analysis target. As a representative bibliometric mapping analytical tool (Cobo, López-Herrera, Herrera-Viedma, & Herrera, 2011), CiteSpace has been used around the world and obtained the most intensive usage in China (Ping, He, & Chen, 2017). We collected journal articles that mentioned CiteSpace by searching several databases including Web of Science and three Chinese databases, and then conducted a content analysis of these articles. We assessed the differences in citation rate of CiteSpace among journals and disciplines. Moreover, we examined whether there were differences between articles which mentioned CiteSpace in the topic fields (including title, keywords and abstract) and articles which did not mention CiteSpace in the topic fields. Overall, we aim to provide a solid foundation for understanding and improving software citation and description, which will enhance software’s status as research outputs and benefit the developers of software.

Data and Methods
The Web of Science (WoS) database and three Chinese full-text journal article databases, including CNKI, Wanfang and CQVIP, were selected as data sources for articles mentioning CiteSpace. “CiteSpace” and “Cite Space” were used as search terms. The time span was limited to 2004-2016 when CiteSpace was developed and largely spread in the world. First, we searched WoS (limited to Science Citation Index Expanded, Social Science Citation Index and Arts & Humanities Citation Index) for all journal articles containing search terms in the topic fields with title, keywords and abstract. Then we retrieved and downloaded the full text of these articles. In total, we obtained a set of 50 full text articles after discarding non-research articles (e.g., editorials, summary of conferences, comments and letters). Second, we searched for search terms in the topic fields (the same as WoS) of CNKI, Wanfang and CQVIP databases, and then limited articles to these published in Chinese core journals. The journals indexed in the Chinese Science Citation Database (CSCD) or Chinese Social Science Citation Index (CSSCI) or Chinese Core Journals List compiled by Peking University (CCJLPU) were regarded as core journals in this study, because CSCD, CSSCI and CCJLPU are widely used for research evaluation in China. We also downloaded the full text of these articles and manually discarded non-research articles to get the second article set. Among the three Chinese full-text journal article databases, only CNKI database provides full text search. To obtain more papers mentioning CiteSpace, we searched CNKI for journal articles containing search terms in the full-text field, and then limited the journals to Chinese core journals. Later, we further refined the results by selecting research articles mentioning search terms only in paper body as the third article set. A total of 1,418 articles (50 from WoS database, 1,368 from Chinese databases) were collected for this study. The search work was ended on 22nd February 2017. Because the three Chinese databases use different subject
classification systems, a journal indexed in more than one of the databases might be assigned to different domains. Since the journals indexed in CSCD has been included in WoS, we assigned each of these journals with articles using CiteSpace to a unique Journal Citation Reports (JCR) subject area based on the subject area that most of its articles belong to.

A content analysis was conducted to investigate the use and citation of CiteSpace in scientific articles. A content-analytic coding scheme for mentions and citations of CiteSpace was created based on the work of Howison & Bullard (2016) and shown in Table 1. Three coders were trained to code these articles. Before they began to code the articles separately, the inter-coder reliability was measured by calculating Fleiss’ kappa (Fleiss, 1971) using ReCal3 (http://dfreelon.org/utils/recalfront/recal3/; Freelon, 2010). The Fleiss’s kappa was scored at 0.951, which is considered as an almost perfect agreement by Landis and Koch (1977).

Finally, Chi-square tests were used to explore the differences of citing behaviour for CiteSpace between English and Chinese Journals. In addition, we also explored the differences between sections containing CiteSpace for citation counts.

Table 1. Coding scheme for mentions and citations of CiteSpace.

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PaperID</td>
<td>The id of a particular paper mentioned CiteSpace.</td>
</tr>
<tr>
<td>Position</td>
<td>A location mentioned CiteSpace, including Title, Keywords, Abstract, Body, Acknowledgement, and Supplement sections.</td>
</tr>
<tr>
<td>Used</td>
<td>A code indicates whether CiteSpace is used in this research.</td>
</tr>
<tr>
<td>Version number</td>
<td>A particular version of CiteSpace.</td>
</tr>
<tr>
<td>Developer</td>
<td>A mention of the developer of CiteSpace.</td>
</tr>
<tr>
<td>URL</td>
<td>A web address of CiteSpace.</td>
</tr>
<tr>
<td>Citation</td>
<td>A code denotes whether CiteSpace is cited in this paper.</td>
</tr>
<tr>
<td>Reference entry</td>
<td>A code denotes an entry linked to CiteSpace in a reference list.</td>
</tr>
<tr>
<td>Cite to publication</td>
<td>A code denotes cite to a particular publication.</td>
</tr>
<tr>
<td>Cite to manual</td>
<td>A code denotes cite to a specific user’s guide or manual.</td>
</tr>
<tr>
<td>Cite to website</td>
<td>A code denotes cite to URL, project name, and other information.</td>
</tr>
</tbody>
</table>

Results

Overview

We identified 45 full text articles using CiteSpace in the 50 articles from WoS and denoted these articles simply as WoS group. It was noting that this study focused on papers using CiteSpace rather than papers mentioning CiteSpace. That is, only articles that were affected by CiteSpace were included for analysis. We allocated 45 articles to countries based on first author’s country of origin. We found that China, USA and Brazil had 30 (67% of 45 articles), 11 (24%) and 4 (9%) articles respectively. The left four articles separately came from Romania, Turkey, England and South Africa. This finding is consistent with a recent study on logs of interactive events, which found that China and USA were the top two countries with the most intensity usage of CiteSpace (Ping, He, & Chen, 2017). We identified 1,020 articles using CiteSpace from the gathered 1,368 Chinese articles. Figure 1 displays the distribution of the 1,020 articles across the 274 journals. We observed that most journals (84% out of 274) published less than five articles using CiteSpace, while four journals published more than 50 articles.
Characteristics of the mentions of CiteSpace

The 45 English journal articles are distributed across 18 JCR disciplines. Computer science, information and library sciences, and cell biology have more articles using CiteSpace than other disciplines. They have, respectively, 18, 4, and 4 articles. The 1,020 Chinese journal articles are distributed across 36 JCR disciplines. We found that most disciplines published less than 20 articles using CiteSpace. Table 2 presents the disciplines containing more than or equal to 20 articles using CiteSpace. It suggests that CiteSpace is more frequently used in information and library sciences, management, and education.

Table 2. The disciplines containing more than or equal to 20 articles using CiteSpace

<table>
<thead>
<tr>
<th>Disciplines</th>
<th>The number of articles using CiteSpace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information and library sciences</td>
<td>380</td>
</tr>
<tr>
<td>Management</td>
<td>158</td>
</tr>
<tr>
<td>Education</td>
<td>153</td>
</tr>
<tr>
<td>Sport sciences</td>
<td>51</td>
</tr>
<tr>
<td>Multidisciplinary social sciences</td>
<td>42</td>
</tr>
<tr>
<td>Geography</td>
<td>29</td>
</tr>
<tr>
<td>Economics</td>
<td>27</td>
</tr>
<tr>
<td>Communication</td>
<td>23</td>
</tr>
<tr>
<td>Environmental sciences</td>
<td>20</td>
</tr>
</tbody>
</table>

The descriptions of software provided in the articles are valuable for identifying and finding software. Among the 1,020 Chinese journal articles, 435 (43%) articles provided the version information of CiteSpace that they used in title, abstract, or article body. In addition, 433 (42%) and 21 (5%) articles, respectively, mentioned the author and the websites of CiteSpace in abstract or article body. It suggests that Chinese authors did not provide sufficient information for finding CiteSpace.

Characteristics of the citation of CiteSpace

Thirty-seven out of 45 (82%) English journal articles provided a formal citation of CiteSpace in the reference list, while 499 out of 1,020 (49%) Chinese journal articles made a formal citation to CiteSpace. The substantial higher citation rate of CiteSpace for the English journal articles might due to all the English journal articles containing CiteSpace in the topic field. In
contrast to the citation rate of software (44%) found in the work of Howison & Bullard (2016),
our citation rate of CiteSpace for Chinese journal articles is slightly higher. This might be
explained by citeable items are provided in the website of CiteSpace.
We then examined the cited references related to CiteSpace. Among the 37 English journal
articles, 34 articles cited a publication related to CiteSpace. 2 provided a citation to a website,
and 1 cited a user’s manual. It shows that authors are more likely to cite publications than
websites and user’s manuals. Twenty-eight articles cite the article “CiteSpace II: detecting
and visualizing emerging trends and transient patterns in scientific literature”, which was
We also found that authors who published in Chinese journals prefer to cite related
publications; 464 articles (93% out of 499 articles) cited a related publication. The top 10
most highly cited items for CiteSpace within Chinese journal articles are shown in Table 3. It
is worth noting that the second most highly cited article is the Chinese version of the first
most highly cited article. Among the top 10 most highly cited items, four publications and the
blog are in Chinese. These Chinese references contribute to the wide use of CiteSpace in
China.

<table>
<thead>
<tr>
<th>Cited item</th>
<th>The number of articles that cited this item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen C. CiteSpace II: 科学文献中新趋势与新动态的识别与可视化[J]. 情报学报, 2009(6): 401-421.</td>
<td>45</td>
</tr>
<tr>
<td>Chaomei Chen’s Homepage. <a href="http://cluster.cis.drexel.edu/~cchen/citespace/">http://cluster.cis.drexel.edu/~cchen/citespace/</a></td>
<td>25</td>
</tr>
<tr>
<td>刘则渊, 陈悦, 侯海燕. 科学知识图谱: 方法与应用[M]. 北京: 人民出版社, 2008.</td>
<td>18</td>
</tr>
<tr>
<td>刘则渊, 陈超美, 胡志刚等. 引文空间分析原理与应用: CiteSpace 实用指南[M]. 北京: 科学出版社, 2014.</td>
<td>9</td>
</tr>
<tr>
<td>Chaomei Chen's blog. <a href="http://blog.sciencenet.cn/u/ChaomeiChen">http://blog.sciencenet.cn/u/ChaomeiChen</a>.</td>
<td>4</td>
</tr>
</tbody>
</table>

We next compared the citation rates of CiteSpace between different article sets. Thirty-seven
out of 45 (82%) English journal articles provided a formal citation of CiteSpace in the
reference list, while four hundred and ninety-nine out of 1,020 (49%) Chinese journal articles
made a formal citation to CiteSpace. The substantial higher citation rate of CiteSpace for the
English journal articles might due to all the English journal articles containing CiteSpace in
the topic field. In contrast to the citation rate of software (44%) found in the work of Howison
& Bullard (2016), our citation rate of CiteSpace for Chinese journal articles is slightly higher.
This might be explained by citeable items are provided in the website of CiteSpace. We then
classified the 1,020 Chinese journal articles into two groups based on whether they contained
CiteSpace in the topic fields or not. The first group, including 528 articles that contained
CiteSpace in the topic fields, was denoted as Chinese topic group. The second one, including
492 articles which did not contain CiteSpace in the topic fields, was denoted as Chinese main
group. Table 4 shows the usage and citation of CiteSpace in the three article groups during the period of 2007-2016. We can find that there is an over upward trend in the number of articles using CiteSpace in Chinese journal article sets. The average citation rate of the three paper sets is 0.82, 0.59 and 0.39 respectively. The citation rate of CiteSpace out of WoS group is higher than that of two Chinese groups. Moreover, the citation rate of CiteSpace out of Chinese topic group is higher than that of Chinese main group.

Table 4. The usage and citation of CiteSpace in the scientific lecture over time.

<table>
<thead>
<tr>
<th>Year</th>
<th>WoS group</th>
<th>Chinese topic group</th>
<th>Chinese main group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A B C</td>
<td>A B C</td>
<td>A B C</td>
</tr>
<tr>
<td>2007</td>
<td>0 0 /</td>
<td>2 2 1.00</td>
<td>1 1 1.00</td>
</tr>
<tr>
<td>2008</td>
<td>2 1 0.50</td>
<td>5 4 0.80</td>
<td>0 0 /</td>
</tr>
<tr>
<td>2009</td>
<td>0 0 /</td>
<td>12 9 0.75</td>
<td>8 4 0.50</td>
</tr>
<tr>
<td>2010</td>
<td>0 0 /</td>
<td>22 15 0.68</td>
<td>15 4 0.27</td>
</tr>
<tr>
<td>2011</td>
<td>4 4 1.00</td>
<td>42 30 0.71</td>
<td>24 16 0.67</td>
</tr>
<tr>
<td>2012</td>
<td>1 1 1.00</td>
<td>55 32 0.58</td>
<td>44 17 0.39</td>
</tr>
<tr>
<td>2013</td>
<td>4 4 1.00</td>
<td>52 38 0.73</td>
<td>75 35 0.47</td>
</tr>
<tr>
<td>2014</td>
<td>11 9 0.82</td>
<td>90 53 0.59</td>
<td>90 44 0.49</td>
</tr>
<tr>
<td>2015</td>
<td>13 10 0.73</td>
<td>104 51 0.49</td>
<td>105 31 0.30</td>
</tr>
<tr>
<td>2016</td>
<td>10 8 0.80</td>
<td>144 75 0.52</td>
<td>130 38 0.29</td>
</tr>
<tr>
<td>Total</td>
<td>45 37 0.82</td>
<td>528 309 0.59</td>
<td>492 190 0.39</td>
</tr>
</tbody>
</table>

Note: A indicates the number of papers using CiteSpace; B indicates the number of papers that provided a formal citation of CiteSpace in the reference list; C indicates the citation rate of CiteSpace which was calculated by B/A.

To further determine whether there is a statistical difference of citing behaviour for CiteSpace between English and Chinese Journals, we employed Chi-square test to compare the citation rate of CiteSpace between WoS group and Chinese topic group using SPSS (SPSS, version 20; IBM Corp., Armonk, NY). A statistically significant difference between the two groups was found (p < 0.05, Table 5). CiteSpace used in English core journal articles is more likely to receive citations than that used in Chinese core journal articles. Moreover, we also assessed the differences between sections containing CiteSpace for citation counts by comparing the citation rate of CiteSpace between Chinese topic group and Chinese main group. We found that Chinese topic group has a lower uncitedness than Chinese main group (p < 0.05, Table 5). We hold that authors put CiteSpace in different sections based on the importance of CiteSpace to their research. That is, authors tend to mention CiteSpace in their articles’ title, keywords and abstract when CiteSpace is very important to their research. While CiteSpace is not very important to their research, they are more likely to mention CiteSpace in article body. Therefore, to some extent, CiteSpace that is important to research is more likely to receive citations.

Table 5. Chi-square tests for comparison of differences in uncitedness.

<table>
<thead>
<tr>
<th>Test</th>
<th>WoS VS. Chinese topic group</th>
<th>Chinese topic VS. Chinese main group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>9.736</td>
<td>40.381</td>
</tr>
<tr>
<td>P</td>
<td>0.020</td>
<td>0.000</td>
</tr>
</tbody>
</table>

We classified the 1,020 Chinese articles, according to the disciplines that their journals belong to. We then calculate the citation rates of CiteSpace for the disciplines which contain more than or equal to 20 articles. As shown in Table 6, the citation rate of CiteSpace varies from one discipline to another ranging from 0.29 (Sport sciences) to 0.57 (Communication). It
demonstrates that software citation practices are far from common within the scientific community. A lot of efforts are needed to examine software attribution and improve software citation practices.

Table 6. The citation rate of CiteSpace of each discipline

<table>
<thead>
<tr>
<th>Disciplines</th>
<th>The citation rate of CiteSpace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>0.57</td>
</tr>
<tr>
<td>Information and library sciences</td>
<td>0.52</td>
</tr>
<tr>
<td>Education</td>
<td>0.52</td>
</tr>
<tr>
<td>Environmental sciences</td>
<td>0.50</td>
</tr>
<tr>
<td>Management</td>
<td>0.49</td>
</tr>
<tr>
<td>Multidisciplinary social sciences</td>
<td>0.38</td>
</tr>
<tr>
<td>Geography</td>
<td>0.34</td>
</tr>
<tr>
<td>Economics</td>
<td>0.26</td>
</tr>
<tr>
<td>Sport sciences</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Conclusion and Future Work

This paper is part of a large effort to examine the attribution, citation and impact of scientific software. In this article, we collected a total of 1,418 full-text articles from English and Chinese databases, and then undertook a context analysis of these articles to how CiteSpace is used and cited among the articles published in English and Chinese core journals. Moreover, we explored the differences of citing behaviour for CiteSpace between English and Chinese Journals, as well as the differences between sections containing CiteSpace for citation counts. Results show the number of Chinese articles using CiteSpace has increased year by year. CiteSpace was more frequently used in information and library sciences, management, and education. Our results also demonstrate that many Chinese authors did not provide sufficient information for identifying CiteSpace in their articles or provide a formal citation of CiteSpace in the reference list. In addition, findings show that there is a significant difference in citation rate of CiteSpace between English and Chinese Journal articles. Findings also demonstrate that articles containing CiteSpace in title, keywords and abstract are more likely to make a formal citation to CiteSpace than those containing CiteSpace in article body.

Our future work includes collecting more English full-text journal articles using CiteSpace to corroborate the findings of this study. We tend to gather still more articles citing the publications related to CiteSpace. We are also interested to explore the difference of citing behaviour for CiteSpace between Chinese and foreign authors. In addition, we would like to investigate how CiteSpace is diffused over time and across disciplines. Research helps to incorporate software as an integral component in research evaluations and scholarly communication.

References


China as Number 1? A Case for China Regaining World Leadership of Science and Technology

A. Basu\textsuperscript{1}, P. Foland\textsuperscript{2}, G. Holdridge\textsuperscript{3} & R. D. Shelton\textsuperscript{4}

\textsuperscript{1}aparnabasu.dr@gmail.com
National Institute of Science Technology and Development Studies
Dr. K. S. Krishnan Marg, New Delhi 110012 (India)

\textsuperscript{2}pfoland14@gmail.com
ITRI, 518 Camp Meade Road, Baltimore, MD 21090 (USA)

\textsuperscript{3}gholdrid.wtec@gmail.com
\textsuperscript{4}shelton@wtec.org
WTEC, 1653 Lititz Pike, #417, Lancaster, PA 17601 (USA)

Abstract
China has a long and proud history of world leadership in science, technology, and industrial innovation, but in the past two centuries it has experienced a period of political and economic instability that has challenged that leadership. However, since its political consolidation in the middle part of the 20\textsuperscript{th} Century and the subsequent introduction of economic reforms in the late 20\textsuperscript{th} Century, China's rise in science has been meteoric. This rise was first detected by the scientometric community through its indicators, but it has now become obvious. Indeed in 2017 the question, "Will China come to lead world science?" is becoming to some, "Does China already lead world science?" This paper tries to make the case that the answer is "yes" (or at least "soon")—but the answer depends on which metrics one considers. China already leads many countries (including the U.S.) in some measures of GDP, scientific paper production, researchers, plus high technology manufacturing and exports, China recently passed the European Union as a whole in R&D investment. Even in some of those indicators where China has not yet taken the lead, reasonable forecasts predict that it soon will. However, there are some indicators where China is still far behind. For example while rising, it still lags the U.S. and EU in citations in Western publications, and will take years to catch up. It may take decades for China to overcome its lag in Nobel prizes, but the Chinese have a very long view of such goals. Here, these quantitative measures are supplemented by qualitative ones from WTEC assessments and by survey results of scientists and the public, which present a more nuanced conclusion. While Chinese leadership may be difficult for Westerners to accept, it can be viewed as China merely regaining its historical position of leadership in science and technology.

Conference Topic
Country-level studies

Introduction
This paper is the latest in a series by the authors reviewing scientometric indicators to illustrate trends in world leadership of science and technology (S&T). The first was presented at the 2003 ISSI conference in Beijing to measure who was leading the race, but the contestants were then only the United States (U.S.) and the European Union (EU)—the People’s Republic of China (PRC) was not yet a contender (Shelton & Holdridge, 2003). But the spectacular progress of China quickly upended the race. At the 2009 ISSI conference in Rio, indicators and forecasts were presented for a race among the U.S., EU, and China (Shelton & Foland, 2009). The last bullet in that presentation was:

- "I predict that, if present trends continue, the PRC will lead the world by 2017."

In 2017 ISSI is back in China, so how did that 2009 prediction hold up? It was based on forecasts of a dozen traditional indicators with data through 2005. The 2009 paper did not include qualitative assessment by peer review or surveys, which can present a different perspective. And some new indicators have recently emerged that ought to be considered.
Why does it matter which nation leads the world? One reason is the same as in sports, a race can inspire everyone to try harder. Nations struggle to find resources to fund their researchers, and viewing progress as a competition can help with motivations. While science provides global benefits, location does matter for economic prosperity and national security.

Since the 1950s, one goal of the U.S. government has been maintaining world leadership in science, mathematics, and engineering, and there is wide acceptance in the United States of the premise that it remains ahead, albeit with growing doubts. The Chinese Mid- to Long-Term S&T Development Plan (2006–2020) set a goal of doubling national R&D investment intensity to 2.5% of GDP by 2020. China is on track to achieve this goal, as confirmed by the most recent update, a plan for 2016–2020 (McLaughlin, 2016). This increasing investment has clearly paid off in China’s rise in science.

Some bibliometricians were alert to China’s advance in indicators like publications: (Moed, 2002), (Jin & Rousseau, 2005), and (Leydesdorff & Zhou, 2005). In 2006 Zhou and Leydesdorff (2006) made the case that China could already be considered a leading nation in science, particularly in nanotechnology.

This paper will update some of these indicators and provide additional ones. Simple extrapolations then allow some insight into what is likely to happen next in the race for world leadership. The paper will also contrast quantitative results with qualitative indicators, such as survey results from recent surveys conducted by the Pew Foundation and by WTEC. To save space, some data is posted in supplementary material at (Basu, et al., 2017).

**Quantitative Indicators**

There are many metrics that could be used to rank nations. Table 1 lists a selection that the authors consider to be relevant here. Most indicators contain the year when China passed the U.S. and EU, or a forecast of when it is likely to do so. Both the current EU28, and the EU27 without the UK, are shown. Each indicator will be discussed briefly in turn.

**Table 1. S&T Indicators. Entries are for the latest year available. The years (or ?) in parentheses are when China passed the U.S. or EU, or is forecast to pass. EU27 is the EU without the UK.**

<table>
<thead>
<tr>
<th>Indicator</th>
<th>U.S.</th>
<th>EU28</th>
<th>EU27</th>
<th>China</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDP</td>
<td>$18.5T</td>
<td>$20.3T</td>
<td>$17.5T</td>
<td>$21.4T</td>
<td>Trillions (10^12), PPP, current dollars in 2016</td>
</tr>
<tr>
<td>Researchers</td>
<td>1.352 T</td>
<td>1.759 T</td>
<td>1.483 T</td>
<td>1.524</td>
<td>Thousands, FTE in 2014</td>
</tr>
<tr>
<td>PhDs in S&amp;E</td>
<td>35,360 T</td>
<td>58,541 T</td>
<td>47,548 T</td>
<td>32,331 2012</td>
<td></td>
</tr>
<tr>
<td>Paper Share</td>
<td>22.0% (2017)</td>
<td>30.0% (2023)</td>
<td>24.7% (2017)</td>
<td>13.0%</td>
<td>Percent of World in 2013</td>
</tr>
<tr>
<td>Patents</td>
<td>53,318 T</td>
<td>51,587 T</td>
<td>45,302 T</td>
<td>25,834 PCT Applications in 2014</td>
<td></td>
</tr>
<tr>
<td>Citations</td>
<td>1.43 T</td>
<td>1.19 T</td>
<td>NA</td>
<td>0.86</td>
<td>Average Relative Citations in 2012</td>
</tr>
<tr>
<td>Nobel Prizes in Science</td>
<td>274 T</td>
<td>332 T</td>
<td>242 T</td>
<td>9</td>
<td>Nationality counted by birth and at time of award.</td>
</tr>
</tbody>
</table>
**GDP**

Gross Domestic Product is a nation's output in goods and services. To compare nations one can either use the prevailing exchange rate or purchasing power parity (PPP) weights to compensate for local prices. In the first case, the U.S. still leads China by $18.5 trillion to $11.8 trillion. However, with the more realistic PPP weighting, China passed the U.S. in 2014 and EU28 in 2015 to lead the world in this overall measure of economic output. The pattern in Fig. 1 is typical of many indicators: China starting far below the U.S. and EU, rising rapidly, so that it passes the others or can be forecast to do so soon. Such extrapolations are linear for the U.S. and EU, but quadratic for China, which provides a better fit to the data. All figures use red triangles for China, blue diamonds for the U.S. and green squares for the EU.

**GERD**

Gross expenditure on research and development (GERD) is the most common indicator of national investments in R&D from both public and private sectors. It has been available for decades from (OECD, 2017). There are some 35 member states, plus 7 others that supply comparable data, called the OECD Group (OECDg) here.

For many years, the U.S. led the world in GERD, with the EU as a whole not far behind. However, China has been rapidly increasing its GERD investment, passing the EU28 in 2014. An extrapolation of these trends suggests that China will probably pass the U.S. to lead the world in 2018.

One GERD variant that is particularly useful is the percentage share of world GERD from a particular country. Shelton (2008) has shown that this is a driver for paper share, which explains China's rapid rise in publications at the West's expense. Fig. 2 shows GERD share based on percentages of the OECDg. Using worldwide data from (UNESCO, 2017) would change these curves only slightly, since the OECDg accounts for more than 90% of the world's GERD. Note: EU27 reflects the Brexit departure of the UK.

![Fig. 1. GDP, $Trillions with PPP weights.](image1)

![Fig. 2. Percent GERD share of OECDg nations.](image2)

**Researchers**

The source for full time equivalent (FTE) researchers is (OECD, 2017). While the Chinese passed the U.S. in 2011, they are not likely to pass the EU28 curve until late in the next decade (Fig. 3). However, if the number from the UK is subtracted from EU28, the Chinese already led the world in this indicator after 2014.

**PhD Degrees in Science and Engineering**

The latest Science and Engineering Indicators volume from the National Science Foundation (NSF) has data on PhD degrees (NSB, 2016). In Fig. 4, the 2012 EU figure comes directly from
this source, but is estimated in other years from data there for six large EU countries. China was on track to pass the U.S. until 2009, when they put in a new emphasis on quality. China did pass the U.S. in 2007 in natural science and engineering degrees only. Also there were about 4000 Chinese S&E PhD graduates of U.S. universities in 2012, many of whom went home. The EU will likely retain the overall lead in this indicator indefinitely, even with the loss of the UK.

![Fig. 3. Full time equivalent researchers. The Chinese figure was adjusted in 2009.](image1)

![Fig. 4. PhD degrees in science and engineering. EU is estimated.](image2)

**Paper Shares**

Shelton and Foland (2009) forecasted scientific paper publications in leading nations based on a model that connected publication shares of the Web of Science (WoS) to R&D investment shares in the OECD group of nations (Shelton, 2008). The data available then ran through 2005 when China's output was small fraction of those of the U.S. and EU. Still the model forecasted in 2009 that China would pass the other two about 2017 to lead the world, based on national plans for investments.

The model is a linear relationship between the $i^{th}$ country's share of R&D investment $g_i$ and its share of (fractional count) papers $p_i$, indexed by WoS. The constant of proportionality $K_i$ is called the relative efficiency, and it varies by country—some are more efficient than others. The equation is most useful for countries where $K_i$ is fairly constant, as it has been for the EU, U.S., and China, since the mid-1990s.

$$p_i = K_i g_i$$

To evaluate the success of the model, one needs to compare publications measured on the same basis as the 2005 data, which came from the NSF Science and Engineering Indicators series. These biennial volumes presented data for decades based on a proprietary subset of the journals in the WoS, using fractional counts. With the latest (NSB, 2016) volume, a switch was made to Scopus (Côté, Roberge, & Archambault 2016), making checks of long term trends more difficult. Shelton and Foland (2017) have extended the NSF series with a program to analyze samples of WoS hit lists. The results were checked against another series from the Fraunhofer ISI institute (Frietsch Helmich & Neuhäusler, 2017).

The results show that the 2009 forecast was reasonably accurate (Fig. 5). It now seems likely that China will indeed pass the U.S. about 2017. Its crossover with the EU28 is likely to be a year or two later (2020) than originally forecast, but because of Brexit, China now seems likely to also pass EU27 about 2017 to lead the world in this key WoS indicator. Similar conclusions
can be reached from the Scopus database (NSB, 2016, Fig. 5-24); China likely passed the U.S. in 2014, and linear extrapolations suggest that it crossed EU27 in 2016 to lead the world in scientific publications.

**Patents**

One patent indicator is the count of patents granted in the domestic patent offices of several countries (WIPO, 2017). In 2011 China's own patent office (SIPO) passed the totals in each of the U.S. Patent and Trademark Office (USPTO), the Japan Patent Office (JPO), and the European Patent Office (EPO). For international comparisons, it is better to use an indicator like the Patent Cooperation Treaty structure, which allows filing of patent applications in multiple countries with a single application. The PCT's increasing popularity makes it a fair way to compare countries in a neutral database (OECD, 2017). Figure 6 shows a forecast that China might take the lead in this measure in about 2021.

**Citations**

Citations are a good measure of quality of individual papers, but need at least two normalizations to be credible as an indicator of national quality. One is by field, because some disciplines tend to cite many more references than others. Another is by year, since citations take many years to accumulate. One measure that does both is the Average Relative Citations (ARC) data available from (NSB, 2016). These are based on Scopus through 2012. The ARC is the average across a geographic region of the relative citations for each publication. The relative citation in turn divides each publication's citation count by the average citation count of all of the same type of publication in that field and year. ARC values are presented for the year of publication, showing the counts of subsequent citations.

A chart is available in the supplementary material. It shows that the average U.S. paper gets about 40% more citations than the worldwide average paper, and that figure is fairly constant with time. China is far behind, but gaining. Of course, most of these citations come from Western journals in English; if domestic Chinese journals were included, the picture would be far different. However, in some targeted fields China has actually passed the U.S. in a related measure, the number of highly cited researchers.

**Highly Cited Researchers (HCRs)**

This metric is based on the 250 top cited researchers in the disciplinary categories defined in the WoS. Data on HCRs were earlier given by Thomson Reuters, but since 2016 the data is published annually by Clarivate Analytics (2016). Table 2 indicates the distribution of highly cited researchers in the top countries in two time blocks; the period 1981-1999, and the year 2016. A decade ago, the U.S. had the largest number of citations and also the largest number of highly cited researchers amongst all countries. More recently, there has been a rapid increase in
the number for China and a decline in the number of HCRs in the U.S. China has risen from rank 18 to fourth place in 2016.

In the earlier period there were just 9 HCRs (0.2% share) in China; the number has grown to 185 (5.7% share) in 2016 (Basu, 2006; Basu and Ghosh, 2017). In total terms China is still far behind the U.S., but in certain strategic areas, the HCR numbers are comparable. In particular, in the subject area of materials science, China has 46 HCRs as compared to 43 in the U.S. (Table 3). China is now actually ahead of the U.S. in HCRs in two strategic areas, engineering in addition to materials science.

Table 2. Highly Cited Researchers (HCR) in top countries by HCR: 1981-1999 and 2016

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HCR (% share)</td>
<td>HCR (% share)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>U.S. 3082 (67.5%)</td>
<td>U.S. 1529 (46.8%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>UK 354 (7.7%)</td>
<td>UK 324 (9.9%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Germany 194 (4.2%)</td>
<td>Germany 187 (5.7%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Japan 176 (3.9%)</td>
<td>China 185 (5.7%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Canada 144 (3.2%)</td>
<td>Australia 115 (3.5%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>France 117 (2.6%)</td>
<td>Canada 102 (3.1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Australia 78 (1.7%)</td>
<td>Netherlands 99 (3.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>......</td>
<td>France 97 (3.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>China (Rank 18) 9 (0.2%)</td>
<td>Switzerland 78 (2.4%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Highly Cited Researchers</th>
<th>U.S.</th>
<th>China</th>
<th>All countries HCR (39 countries)</th>
<th>U.S.</th>
<th>China</th>
<th>All countries HCR (53 countries)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCR (All fields)</td>
<td>3082 (67.5%)</td>
<td>9 (0.2%)</td>
<td>4569 (46.8%)</td>
<td>1529 (46.8%)</td>
<td>185 (5.7%)</td>
<td>3265 (5.7%)</td>
</tr>
<tr>
<td>Rank by HCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Fields</td>
<td>1</td>
<td>18</td>
<td>na</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Materials Science</td>
<td>1</td>
<td>&gt;10</td>
<td>na</td>
<td>2 (43)</td>
<td>1 (46)</td>
<td>152 (43)</td>
</tr>
<tr>
<td>Engineering</td>
<td>1</td>
<td>9</td>
<td>na</td>
<td>2 (25)</td>
<td>1 (37)</td>
<td>145 (25)</td>
</tr>
<tr>
<td>Computer Science</td>
<td>1</td>
<td>9</td>
<td>na</td>
<td>1 (53)</td>
<td>2 (13)</td>
<td>126 (53)</td>
</tr>
<tr>
<td>Chemistry</td>
<td>1</td>
<td>&gt;10</td>
<td>na</td>
<td>1 (71)</td>
<td>2 (7)</td>
<td>215 (71)</td>
</tr>
<tr>
<td>Geosciences</td>
<td>1</td>
<td>&gt;10</td>
<td>na</td>
<td>1 (60)</td>
<td>3 (12)</td>
<td>149 (60)</td>
</tr>
</tbody>
</table>

High-Technology Manufacturing and Exports

A scientometric indicator that can be viewed as the bottom line of the innovation cycle is the performance of high-technology (HT) industries. Data on HT exports have been complied on a cash basis for decades by the OECD (2017). Figure 7 shows that China took the world lead in this measure in 2005 (NSB, 2016). However, these export measures of industrial output do not capture the nuances of where manufacturing really takes place. Recently a new dataset has been developed by the OECD and the World Trade Organization for manufacturing output on a value-added basis, which avoids double-counting of imported components. This more accurate
data, as summarized in (NSB, 2016), allows development of much-improved models that tie these key outputs to inputs like R&D investment (Foland, Fadel & Shelton, 2015). Figure 8 shows some national time series for this measure. Percentages are based on current dollars at the prevailing exchange rate. China likely took the world lead by 2016.

Nobel Prizes

Nobel Prizes are the gold standard metric for scientific excellence (see Table 1). (There are also prizes for literature, peace, and economics, but they are not included here.) Scientists often migrate during their career, which makes counting their nationality somewhat inexact; the approach here is to credit both the country of birth and country at the time of award (Nobel, 2017). Western countries have accumulated a huge head start since the prizes were first awarded in 1907. It will take China a very long time to catch up, but it could compete sooner in prizes in recent years.

Qualitative Indicators

Anecdotal Information from WTEC International Studies

WTEC has conducted over 70 international technology assessment studies since 1989 on behalf of the NSF and other U.S. government agencies, each focused on a field of science or engineering. Most involved site visits by a panel of U.S. experts to leading laboratories, factories, and funding agencies. Because of the expense of study tours, the study topics and the sites to be visited are chosen with extreme care. Prior to 2000, none of the panels visited China, which was an indication that it was not yet considered to be among the leading nations in the topics at hand.

However, since 2000, ten WTEC study panels have traveled to China and conferred with over 400 Chinese researchers. Some of these meetings were during visits to leading labs; some were during workshops organized by WTEC in China. The fact that visits to the PRC were deemed essential to the more recent WTEC studies is another metric of China’s rising status as an international power in S&T.

While space precludes discussion of findings from all these studies, one example can illustrate the trends seen by WTEC panels. This is supercomputer technology, which has been explicitly a race for world leadership since at least 1980. In 2004 WTEC sent a delegation to Japan, which had briefly taken the lead in the race. At the time, China was hardly present in the "Top500" list of supercomputers in operation, much less in development of its own models. However, its progress was rapid, and by 2009 another WTEC delegation evaluated software for modeling and simulation at 59 sites worldwide, including nine in China, where the panel was impressed with their progress. Fast forward to 2016, when China now clearly leads the world.
in supercomputers, with not only the fastest ones, which are also made with indigenous components, but also with as many installed machines on the top500 list as the U.S. (Top500, 2016). While the U.S. and EU still have plans to compete for leadership, they seem to be falling further behind.

An analysis of the findings from other WTEC studies yields a variety of qualitative indicators of (anecdotal information on) China’s rising S&T leadership. The major conclusions are: (1) China is following a long-term S&T investment strategy, including a strong focus on topics of relevance to technology-based economic development and national security. (2) Similarly, China’s long-term strategy for developing its S&T education and high-technology workforce has been extremely effective. (3) As a result, there is an increasing number of specific researchers and institutes that were deemed by the visiting WTEC panelists to be doing world-class work. A synopsis is available in the Supplementary Information (Basu, et al., 2017)

Evidence from Expert Surveys
1. Pew Foundation Survey
The Pew Foundation has periodically surveyed members of the public and members of the American Association for the Advancement of Science (AAAS) for their opinions on the status of U.S. S&T (Funk & Rainey, 2015). While diverse, many AAAS members are life scientists. Some of the findings from the Pew survey are:
- Less than half of AAAS members surveyed believed the U.S. leads the world in scientific achievements. It was 45% in 2014, down from 49% in 2009.
- Responses on this question from the general public are far lower: 15% in 2014, vs. 17% in 2009.
- 29% of AAAS respondents believed the U.S. is the leader in industrial R&D innovation.
- 40% of AAAS respondents believed the U.S. leads in basic research, and 46% believe it leads in doctoral training in S&T.

Pew’s synopsis of the most important findings from the most recent 2014 survey are:
- Both the public and scientists see U.S. scientific achievements in a positive light. But they are critical of K-12 STEM education.
- Scientists are also still largely positive, but less upbeat than five years ago [2009].

2. WTEC Surveys of Scientometrics Authors and WTEC Panelists
WTEC recently conducted a survey to complement the Pew survey with expert opinions on world leadership in S&T. Two samples were queried, each with N = 100 valid responses: (1) authors of recent papers in Scientometrics and (2) experts who have served on WTEC international assessment panels. Both groups can be said to have special knowledge of national standings in S&T. Questions and responses from both samples are available at (Basu, et al., 2017). Two key questions concerned current overall leadership in S&T, and the respondents’ opinions on the status 20 years from now. A chart summarizing the results is shown in Fig. 11. Both groups rated the U.S. as currently leading the world in S&T. However, both groups also agreed that the U.S. position will decline in the next 20 years. About half of both samples thought that China might assume world leadership in 20 years.

U.S. K-12 education in S&T was rated by the WTEC survey respondents as inferior to China's, consistent with international studies like (PISA, 2017). The current K-12 student cohort will be a key part of the scientific workforce during the next 20 years. While there may be systemic problems with the U.S. K-12 S&T education system on average across the whole country, in most countries S&T leaders are likely to come from elite K-12 institutions. Thus, the notion
that its K-12 S&T education in general is deficient does not necessarily mean that the U.S. will not produce an adequate number of world-class S&T leaders for 20 years hence.

More immediately, Fig. 12 shows that, like the AAAS members, many respondents believe the U.S. currently leads in doctoral education. However, if the U.S. is no longer able to attract doctoral program graduates to remain, current advantages in the U.S. doctoral programs may not be translated into future S&T leadership.

Impact of Decreased Accessibility of U.S Science and Engineering Education.

One consideration with respect to science education that may be an important factor in these comparisons between countries is the accessibility of science and engineering education. Raw talent in S&T is not restricted to those of wealthy means. Therefore societies that expect to lead the world in S&T need to nurture those individuals who have that talent, regardless of their economic backgrounds. Because success in science and engineering education requires not only high intelligence, but also hard work, it may even be more likely that gifted individuals of limited means with a strong work ethic, are more likely to excel in these fields than those who have grown up in a life of ease.

A couple of decades ago, technical education in the U.S. was far more affordable than it is today. In the mid-20th Century it was possible for Americans from low-income backgrounds to excel at the K-12 education level, qualify for top-flight U.S. undergraduate programs, and work hard to pay their way through these programs. Then, having excelled at the undergraduate level, these Americans qualified for graduate programs supported by the U.S. Government, and went on to become leaders of late-20th Century science and engineering. And those leaders were accompanied by many of the world's best from abroad. It is not clear that this continues to be the case: College education is increasingly out of the price range of Americans of modest means. Real costs of attending universities have doubled and doubled again in the last 50 years or so, while salaries have stagnated, rising only 10% or so in real terms. And recently there is doubt about whether those talented immigrants will continue to come to the U.S.

Altmetrics Results

Altmetrics are statistics obtained from social media and online resources counting mentions, downloads, reads, tweets, recommendations, Facebook entries and so on, about research papers. They indicate informal ways of noting the relevance of the content, and its importance to the reader. Some of these can culminate in citations, but usually that is a small percentage. The advantage of using altmetrics instead of citations has been discussed by many; see (Fairclough,
2015) and references therein. For one, they give an indication of potential usefulness of a scholarly work much earlier than is possible for citations. Citations take a long time to accumulate, the time delay starting from the point of submission of a paper, through the peer review process, publication delays, being seen by the reader, and finally, the time taken to incorporate the work as a reference. In the case of altmetrics, the paper may be available early, say from an Open Access repository (even prior to publication), and information about it can be immediately tweeted or communicated through blogs, etc. The measure of attention, or altmetric score, given to a paper depends on the frequency of tweets about the paper, as well as the number of downloads, reads etc., on various social media platforms. By this method, one can utilize the enormous data generated on social media based on numerous individual actions, and use it to generate some crowd-based insight.

For country-level estimates, countries differ in terms of altmetric scores, but this is not only due to the quality of their papers. Some countries like China have restrictions on using Twitter and Facebook. Therefore a Chinese paper would lose out on contributions to its altmetric score from readers in China. One of the favoured sites for obtaining altmetric data is Mendeley, a reference manager used by scientists, typically for reference storing and sharing. Mendeley has been examined for reader counts and found to be a useful source of early impact information. It also correlates well with citations (Thelwall, 2015; Wang, 2016). However, some biases such as national dominance of some countries in the use of Web-based social media, and (expected) dominance of junior colleagues over senior (due to greater facility of the young in the use of new technology), are factors that need be taken into account. Thus proper interpretation and international comparisons may be difficult for Web indicators because there are basic national differences in the extent of use of the Web, and as a result there can be large differences in the uptake of the social websites (Thelwall, 2015).

In 2016, results of a study were released by Altmetric.com that looked at 17 million mentions of 2.7 million articles published the previous year. The mentions included news stories, blog posts, tweets, Facebook posts, Reddit posts, articles in F1000 (a manually curated list of excellent papers), Wikipedia citations, and mentions on Mendeley. Researchers from over 440 institutions contributed to the papers that made up the Top-100 list for 2016, with authors spread out over 42 countries. Of the top 100 papers by altmetric score are five papers from China at ranks 3, 43,64, 84, and 86 (Altmetric, 2017).

<table>
<thead>
<tr>
<th>Ser</th>
<th>Country</th>
<th>Papers</th>
<th>Altmetric Score</th>
<th>No. of Authors</th>
<th>Altmetric Score/ Paper</th>
<th>Altmetric Score/ Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>EU</td>
<td>73</td>
<td>173013</td>
<td>772</td>
<td>2370</td>
<td>224.1</td>
</tr>
<tr>
<td>2</td>
<td>China</td>
<td>5*</td>
<td>12007</td>
<td>17</td>
<td>2401</td>
<td>706.3</td>
</tr>
<tr>
<td>3</td>
<td>US</td>
<td>74</td>
<td>169456</td>
<td>847</td>
<td>2290</td>
<td>200.1</td>
</tr>
</tbody>
</table>

*One paper was later retracted*

Table 4 shows that only a few Chinese papers got enough attention to be in the top 100 papers by the Altmetric Attention Score. The U.S. and EU have approximately the same number of papers in the list (many of them are common due to collaboration). The Altmetric score per paper is approximately the same for the EU, U.S., and China, but the score per author is considerably higher for China. This is largely due to collaborative papers.
Conclusions
This paper has presented a variety of measures of national science leadership to examine the case for China's resuming its historical leadership of science and technology. Generally, quantitative indicators show that China has made enormous progress in S&T. Some of these indicators show that China is already in the lead, and others forecast that it soon will be.

Qualitative indicators like survey results are less sanguine about China's position; few respondents think that China now leads the world or soon will. However, they believe that China is improving its position and many think that it might pass the U.S. in 20 years or so.

Of course, there are many benefits of science and technology, regardless of where the R&D is done. Science can lead to better healthcare, cleaner air and water, solutions of problems like global warming, improved communications that allow more extensive cooperation and collaboration, and many others. Most of these benefits can accrue to everyone, regardless of their nationality.

References
Moed, H. F. (2002), Measuring China’s research performance using the Science Citation Index. Scientometrics 53: 281-296.
A Peer Review Method Based on Identify Experts’ Guanxi

Yuan Junpeng¹ Wang Xianhui ² Gao Jiping¹ Su Cheng¹

¹junpengyuan@gmail.com
Institute of Scientific and Technical Information of China, Beijing (China)

²165663818@qq.com
Research Center for Shandong Science and Technology Statistics, Jinan (China)

Abstract
In Chinese society, guanxi (interpersonal relationships) are involved in everyday social interactions. In the processing of peer review, interpersonal relationships between experts and person to be reviewed often interfere the objectively assessing of evaluation, leading to the review invalid or injustice. According to the requirement of different review and the setting of review threshold, this paper proposes a bold scientific evaluation idea for reasonable quantitative analysis of these relationship. The selection of assessment experts refers to quantitative evaluation of interpersonal relationships and review threshold. In the reviewing process, the selection of expert is combined with a review of appraisal results to make the evaluation result more credible. To verify the feasibility of this study, the instance will be reviewed in the experts’ interpersonal relationships, which will be refined and applied in practical assessment.

Conference Topic:
Science policy and research assessment

Introduction
As one of the highest credibility and most widely used evaluation method in research assessment system, peer review plays an important role in scientific research resource allocation. However, in the processing of project review, guanxi (interpersonal relationships) between experts and person to be reviewed often interfere the objectively assessing of evaluation, leading to the review invalid or injustice. In the reviewing process, expert’s avoidance principle is always applied to reduce the impact of guanxi of experts, making sure that the review is objective and valid. In practically review, with a signal avoidance principle, the review could face dilemmas like the less choice and harder selection of experts. Therefore, how to face the various interpersonal relationships of experts, how to apply the avoidance system flexibly, and how to ensure the justice and authority of the evaluation to the maximum extent are all worth studying.

Kostoff (1997) defines the system of peer review as a comprehensive one, which requires assessment from peer experts of various types rather than from a certain domain. Fisher (1999) focus on the influence of moderators during the review process. Bloom (1999) deems that a rational setting process is the key to obtain high-quality evaluation results. Callaham (2001) deeply studied the phenomenon of conservatism in the process of peer review. Bornmann (2011) investigate scientific peer review. Gleditsch and Nordas et al. (2017) introduced peer review system of Journal of Peace Research. Using public information about the identities of 9000 editors and 43000 reviewers from the Frontiers series of journals, Helmer and Schottdorf et al. (2017) show that women are underrepresented in the peer-review process, that editors of both genders operate with substantial same-gender preference (homophily), and that the mechanisms of this homophily are gender-dependent. Moustafa (2017) investigate either open peer-review fully or blind it wholly. Price and Flach (2017) set up a computational support for academic peer review.

Some scholars have also studied the guanxi in China. Wei and Liu et al. (2010) examined the role of political skill in the dynamics of supervisor-subordinate relationship in Chinese firms,
found that supervisor-subordinate guanxi mediated the relationship between political skill and career development of the subordinates. Gao and Knight et al. (2012) aims to identify critical aspects of Chinese-Western intercultural guanxi relationships that have largely been ignored as a domain for study in international business and industrial marketing, and to suggest a way forward. Chen and Chen et al. (2013) review research on Chinese guanxi and social networking in the past twenty years and identify the major perspectives, theories, and methodologies used in guanxi research at micro and macro levels. Li and Qiu et al. (2016) conducted a quasi-experiment to investigate the effects of two significant cultural values, face and guanxi, on Chinese customers' behavioral responses to hospitality service failures. Zheng and Zhang et al. (2017) study the secrecy capacity of large-scale wireless networks with social relationships. Firstly, this paper summarized and analysed the genetically, industrial, geographical relations between experts and person to be reviewed and weight the possibly influences caused by these social relationships. Secondly, according to the requirement of different review and the setting of review threshold, this paper proposes a bold science evaluation idea for reasonable quantitative analysis of these relationships. The selection of assessment experts is reference to quantitative evaluation of social relations and review threshold. In the reviewing, the chosen of expert is combined with a review of appraisal results, so as to make the evaluation result more credibility. Thirdly, to verify the feasibility of this study, the instance will be reviewed in the experts’ social relations of Shandong Province which will be refined and applied in practical assessment. Chose Shandong Province as an example, for two reasons: firstly, the science and technology development of Shandong Province has synchronized with China’s overall development, secondly, as a province, if avoidance expert in some project review, there will no experts are available. This improved comprehensive evaluation method, not only provides the corrected evaluation of social relations affected by evaluation experts, but also reduces the subjective influence of social relations through different term weighting. This pilot trial have many positive effects, such as the effective and impartiality evaluating in science and technology consulting, and the optimal promotion in technological resources allocation.

Research method

The influence of various interpersonal relationships on peer review

Every expert involved in peer review, who is "social man", has his own interpersonal relationships network. There are a variety of interpersonal relationships between the experts. It is obvious that their social networks are mainly divided into three aspects: kinship, business relationship and friendship relationship. These three kinds of interpersonal relationships affect the peer review to varying degrees.

In this paper, relative relationship is divided into two types of interpersonal relationships: family and relatives. The former mainly includes spouses, parents, children, brothers and other immediate family members. This type of relationship has the most direct impact on peer review, and should be taken a strict avoidance action in the actual review generally; The latter refers to the relatively distant relatives of kinship, and there is no need to strictly avoid it in the actual assessment.

The business relationship, as an important part of the interpersonal relationship network, consists of the most abundant interpersonal relationships, including leaders, colleagues, academic circles, students, teachers, students and so on. In the actual review, these relationships are usually avoided. But there will be many drawbacks if we avoid it absolutely.

Friendship is a relationship based on the common hometown, region, community, hobbies, views and other factors, including the fellow-townsmen, friends, e-pals and so on. Friendship relationship is difficult to obtain, and the definition of it is also a little subjective. In many cases,
we have to ignore the impact of these relationships on the review when finding the interpersonal relationships of experts involved in peer review. However, most private relations (such as genetic relations) involve the privacy issues of experts and can’t be obtained by the current scientific research from the channels such as citizenship. Therefore, this paper mainly chooses expert relationships, which can be obtained from the public information and related to the academic work of the experts. This includes affiliated organization relationships, such as the same workplace, interdepartmental relationships, as well as scientific research collaboration relationships, such as project cooperation, academic exchanges, and collaborative researches and so on.

**The acquisition of interpersonal relationships**

The access to the various relationships included in this paper is shown in Figure 1.

---

**Figure 1. Access to the various relationships included in this paper**

a) All data of the information about teachers, classmates and students should be cleaned according to the expert’s personal resume, so that the accuracy about the information can be guaranteed.

b) This paper obtain the information about the expert’s Chinese papers’ co-authors through visiting the Chinese Knowledge Website (http://cnki.net) and combining the information with the expert’s personal resume, and obtain the information about the expert’s English papers’ co-
authors through Web of Science. The information about both Chinese and English papers are then compared and checked to delete the duplicate data. And finally it draws a list of the expert’s collaborators.

Develop a self-programming, construct an author’s cooperative matrix by processing, cleaning, counting, and analysing the downloaded data of paper-cooperation with the self-programming. Normalize the data, that is, give the statistical frequency distribution of the probability distribution analysis. The linear function conversion is as follows:

\[ Y = \frac{(X - \text{MinValue})}{\text{MaxValue} - \text{MinValue}} \]

in which, \( X \) is the frequency of paper-cooperation, \( Y \) is the assignment of the normalized frequency of paper-cooperation, \( \text{MaxValue} \) and \( \text{MinValue} \) are the maximum and minimum values of the cooperation frequency respectively.

Build the author’s cooperative matrix into the social networking analysis software UCINET, use the built-in Netdraw to carry out the central potential analysis, and draw the network of the expert’s paper-cooperation.

c) Sort out the different relationship between the expert and others, construct the corresponding relationship matrix, and build the constructed matrix into the social networking analysis software UCINET. Use the built-in Netdraw to carry out the central potential analysis, and draw the interpersonal relationships network, in which the value of the interpersonal relationships is represented by the Width of the relationship line.

**Weight setting**

Taking the operability and credibility of the results into account, this paper adopts the method of expert evaluation to set the weight, which is, gather experts’ reviews on the questions to be predicted and give them feedback after organizing, summarizing and collecting all the reviews, and then repeat the procedure until a consensus is reached. And during the weight-setting process, experts are not allowed to discuss with each other or make some linkages, but communicate with the investigators, and give their feedback and revision. Finally, the investigators collect the almost consistent views given by experts as the final result of the expert investigation.

A specialized weight-setting team of 10 members are invited to give quantitative assignment of the various interpersonal relationships in this paper. The formula is:

\[ M_j = \frac{1}{m_j} \sum_{i=1}^{m_j} C_{ij} \]

In which, \( M_j \) represents the arithmetical mean of the interpersonal relationships of the group members to the \( j \) \((j = 1, 2, \cdots, n)\), \( m_j \) represents the number of experts in assessing interpersonal relationships of \( j \), \( C_{ij} \) represents the score that the \( i \) \((i = 1, 2, \cdots, m)\) group member gives on the \( j \) interpersonal relationships.

The value of \( M_j \) varies from 0 to 1. The larger the value of \( M_j \) is, the greater the relative importance of the scheme is.

**Review threshold setting**

A case-by-case review system should be in place because different reviews have different impact on the interpersonal relationships of experts. For example, paper review and reward review differentiate greatly in the principle for setting avoidance on interpersonal relationships. Thus, in the actual reviewing process, the method of expert evaluation is applied to set the review threshold according to the varied requirements for evaluation.

A specialized threshold-setting team of 10 members are invited to give quantitative assessment, fully considering the affecting factors, such as the importance of the review, the project approval rate, the fitting degree of industrial policies and so on. The formula is:


In which, $T_j$ represents the arithmetical mean of the review threshold set by the group members to $j$ ($j = 1, 2, \cdots, n$), $t_j$ represents the number of experts in setting review threshold for $j$, $C_{ij}$ represents the score that the $i$ ($i = 1, 2, \cdots, m$) group member gives on $j$. The value of $T_j$ varies from 0 to 1. The larger the value of $T_j$ is, the more qualified relationship the review has.

Peer Review Method for Interpersonal relationships

After complicated interpersonal relationships of experts are extracted in different ways, different expert interpersonal relationships are given different weights. These relationship analyses are identified and quantified reasonably. At the same time, the review thresholds are set according to different requirements. The selection of assessment experts is based on the quantitative evaluation of interpersonal relationships and the different review thresholds. In the review, we analyse and identify effective interpersonal relationships, weigh the possible impact of them on the review, establish the calculation models of correlations.

In the review, in addition to those who are taken the initiative to ask to avoid by the reviewers, the others to be reviewed need to be compared in the relationship lists. If the value of the relationship is higher than the review threshold, it is necessary to perform the avoidance processing or weight the evaluation result. If the value of the relationship is lower than the review threshold, we can allow them to participate in the review, or weigh the evaluation result.

The overall process is shown in Figure 2:

![Figure 2. Peer review process for interpersonal relationships](image)

Empirical Research

In this section, in order to verify the feasibility of this idea, we will refine the interpersonal relationships of experts and apply to the actual reviews in the following examples. (In order to respect the experts' personal privacy, letters are used instead of the experts' actual names.)
By using the method mentioned above, we can get the information of the colleagues, instructors, classmates, students, and the co-authors of the expert Z. Then we can also refer to the weight value from the expert evaluation method and get the quantitative value of the expert Z’s interpersonal relationships by using simple calculation program designed by ourselves. The partial list is as follows:

<table>
<thead>
<tr>
<th>Relationship types</th>
<th>The social relations of expert Z</th>
<th>Department colleague</th>
<th>Other colleague</th>
<th>Instructor</th>
<th>Classmate</th>
<th>Student</th>
<th>Co-author</th>
<th>The overall quantitative value of interpersonal relationships</th>
</tr>
</thead>
<tbody>
<tr>
<td>Related person R1</td>
<td>0.04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Related person R2</td>
<td>0</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td>0.0069</td>
<td>0.2069</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Related person R3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>0.0034</td>
<td>0.2034</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Related person R4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>0.0069</td>
<td>0.2069</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Related person R5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0034</td>
<td>0.0034</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Related person R6</td>
<td>0.04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Related person RN</td>
<td>......</td>
<td>......</td>
<td>......</td>
<td>......</td>
<td>......</td>
<td>......</td>
<td>......</td>
<td></td>
</tr>
</tbody>
</table>

Using the same method, the quantitative value of interpersonal relationships of the remaining experts (expert P, expert Y, expert L, expert W) in the evaluation group can be obtained.

In actual review, we first extract the information of the declarant, the technical director, and the corporation manager of a project, and compare it with the interpersonal relationships personnel of the experts. Then we select the experts according to the review threshold. Assuming that the panellists have set a review threshold to 0.2038, then select the people related to the expert with the expert relationship assignment greater than or equal to 0.2038 to form a list.

In the review, in addition to those who are taken the initiative to ask to avoid by the reviewers, the others to be reviewed need to be compared in the relationship lists. If the value of the relationship is higher than 0.2038, it is necessary to perform the avoidance processing, and that reviewer will have no qualification to review the project. If the value of the relationship is lower than 0.2038, we can allow them to participate in the review, however, it is also necessary to weigh the evaluation result according to the actual requirements of the review if they do have a relationship.

For example, in an annual review of Shandong Province that expert Z participate, the information of applicants including declarants of a certain project, technical directors and managers of the corporations is as follows:

YTB, ZRX, RGD, LZBin, LZBo, TYX, DQ, LJE, PSB, LMH, ZLJ, XYN, YCL, WY, WZG, ZX, QCR, LM, WYu, SDF, GHF, WXbo, WWH, MXP, PRJ.

There are 5 experts in the list of recommended members of the review: expert Z, expert P, expert Y, expert L, expert W.

In the expert Z relationship list, the personnel and the assignment of personnel relations are: LZBin 0.3.

In the expert P relationship list, the personnel and the assignment of personnel relations are: QCR 0.04.

In the expert Y relationship list, the personnel and the assignment of personnel relations are: WY 0.1677, WYu 0.0154.

In the expert L relationship list, the personnel and the assignment of personnel relations are: ZLJ 0.1.
In the expert M relationship list, the personnel and the assignment of personnel relations are: ZRX 0.0056, PSB 0.2056.

It’s easy to distinguish that experts whose relationship assignment are greater than the review threshold of 0.2038 include expert Z and expert W, and that means it is necessary to be judged by the evaluation committee or the management department whether they should be avoided or given the chance to participate in the review. However, the evaluation result should be weighed at last.

The scores of the project given by five experts (expert Z, expert P, expert Y, expert L, expert W) are: 90, 84, 79, 89 and 91.

The project final score got by using the original evaluation method is the average of scores from five experts, 86.6.

If we avoid the expert Z and expert W who own the relationship assignment over 0.2038, the scores given by the other experts are 84, 79, 89, the average of the project “Container terminal production management and control system V4.0” will be 84.

If we don’t avoid them and weigh the scores given by two experts (score/relationship assignment *0.2038), the final score will be: 61.14, 84, 79, 89, 90.2, and the average of the project will be 80.668.

In this way, we can make the score from subjective evaluation more objective, and be able to respond to the public questioning fairly and reasonably.

**Conclusion**

This improved comprehensive evaluation method weighs the possible impact of different interpersonal relationships on the assessment. During the appraisal process, it can analyse and identify the various relationships between the reviewer and the one to be reviewed, as well as quantify them reasonably according to different assessment requirements in order to reduce the subjective impact of interpersonal relationships. It not only considers the influence of the interpersonal relationships of the reviewers, but also reduces the subjective influence of interpersonal relationships by setting different weight. This idea may well have a positive impact on the effectiveness and justice of all kinds of scientific and technological advice and evaluation, promote the optimal allocation of scientific and technological resources and improve the management level of science and technology. This is a meaningful scientific evaluation conception as well as attempt. It fits actual requirements of the development of the current assessment and can be applied to a lot of areas.

This paper focuses on how to reduce the subjective influence of experts’ interpersonal relationships, and puts forward the innovative idea of weighting the evaluation results according to the close degree of interpersonal relationships, or just strictly avoiding them, according to the different relations, which makes the complex comprehensive evaluation more scientific and credible. This improved comprehensive evaluation method can be applied in many fields, and it is suitable for the comprehensive evaluation problem with both quantitative and qualitative evaluation opinions.

The science and technology project evaluation system based on interpersonal relationships is very conducive to scientific research management departments to select peer experts for different projects. Varieties of interpersonal relationships can be visualized in the management system, and the social relation between the commentator and the person reviewed is at a glance. Moreover, the quantitative value of different interpersonal relationships reduces the subjective impact of expert scoring, and thus contributes to a more fair and reasonable, more credible evaluation result.

However, there are a lot of things that need to be improved, such as how to set the weight value, to get rid of the subjectivity. How to set the threshold of the evaluation to be more credible, in the establishment of the network of experts’ guanxi, the problem of the name disambiguation is
more difficult to check, many hidden social guanxi cannot be measured, there are also guanxi
that can be quantified through various channels. The costs involved in a large scale
implementation of the proposed approach seem fairly substantial unless it is possible to
automatize a substantial part of the work. Except for large projects the costs may not be
outweighed by benefits that might be achieved through less cost intensive approaches. Further
research is needed to determine whether this review can be operable in a larger review. The
case presented suggests the involvement of strongly related experts can have a depressing effect
on review scores as well. Further assessments may provide insight in the extent that the apparent
problem is actually strongly prevalent in the Chinese context and/or whether it can be resolved
by the proposed approach.

Acknowledgments
This paper was supported by a grant (No: 71473236) from the National Natural Science
Foundation of China (NSFC)

References
papers by peer review. Journal of the American Medical Association, 272, 143-146.
Callaham, M.L. & Wears, R.L. (2001). Positive-outcome bias and other limitations in the outcome of
research abstracts submitted to a scientific meeting. Journal of the American Medical Association,
280, 254-257.
45, 199-245
Dermatology, 56, 86-88.
from Artificial Intelligence. Communications of the ACM, 60, 70-79.
Wei, L. & Liu, J., et al. (2010). Political Skill, Supervisor-Subordinate Guanxi and Career Prospects in
The small-world phenomenon of microblog communication networks in China: an empirical study based on complex network analysis

Xin Jiang¹, Haiqun Ma² and Dezhuang Wang³

¹ jx_happy@sohu.com
Heilongjiang University, Harbin (China)

² mahaiqun@163.com
Heilongjiang University, Harbin (China)

³ wangdezhuang@163.com
Heilongjiang University, Harbin (China)

Abstract
The aim of this study is to demonstrate that microblog communication networks in China present the small-world property based on the theory and method of Complex Network Analysis (CNA). This study selects Tencent Weibo to conduct empirical study, which is one of the most popular microblogs in China. An actual communication network of celebrity users on Tencent Weibo is constructed using the method of snowball sampling, then it’s proved to present the small-world property by calculating its clustering coefficient and average path distance which are the most important two parameters used for identifying a small-world network. The pivotal users of communication network are determined by calculating the node betweenness centrality of each node, which frequently act as the “opinion leaders” in the process of information dissemination. The pivotal relations between users in the communication network are determined by calculating the edge betweenness centrality of each edge, which usually originate from one of the pivotal nodes and may act as the “long-range edges” in the whole communication network. This study provides theoretical and practical foundation for related research on information dissemination on microblogs with the theory and method of complex network analysis.

Conference Topic
Social network analysis

Introduction
The earliest research on small-world phenomenon was published by Milgram (1967) who designed the famous “chain letters” experiment and found that any person in the United States usually could be contacted by any other person passing through six persons on average, which is popular known as the inference of “six degrees of separation”. Watts and Strogatz (1998) found that the small-world phenomenon exists in many biological, technological and social networks which lie between the extremes of completely regular and completely random. Newman and Watts (1999) put forward a small-world network model which displays a normal continuous phase of transition between regular networks and seldom networks in the limit of large system size. Barabási and Albert (1999) further proposed that many large networks have a common property that the vertex connectivities of them follow the scale-free power-law distribution.

The small-world phenomenon has been found to exist in such many disciplines as materials science (Shi and Guan 2016), engineering (Song et al. 2014), computer science (Eppstein et al. 2013), management (Chen and Jaw 2014; Opsahl et al. 2017), economics (Gallo 2012; Sheehan and Young 2015), business (Feyen 2015), psychology (De-Marcos et al. 2016; Sankowska and Siudak 2016) and social science (Seltzer and Smirnov 2015). In the discipline of Library and Information Science (LIS), Hung and Wang (2010) examined and characterized the small world phenomenon in the patent citations network by analyzing the data of radio frequency identification (RFID) patents. Zhu et al. (2013) found that the keyword co-occurrence network of the LIS discipline presented the small-world property based on the keyword data extracted from Scopus. Zhang et al. (2014) examined the evolution...
of small world network and its impact on patent productivity in China based on the patent co-authorship data from State Intellectual Property Office.

In particular, Watts and Strogatz (1998) put forward a random rewiring procedure for interpolating between a regular network and a random network without altering the number of nodes or edges in the graph. The random rewiring procedure is as follows: starts from a regular network of \( N \) nodes with each node connected to its \( K/2 \) nearest neighbors at both sides (\( K \) is an even number), then randomly rewires each of these edges with probability \( p \), and continue this process with self-loop or duplicate edges forbidden. There are \( pNK/2 \) long-range edges are introduced in this rewiring process which connect several non-adjacent nodes. For \( p=0 \), the original graph is a regular network; as \( p \) increases, the graph becomes increasingly disordered; until \( p=1 \), the graph finally changes into a random network. For intermediate values of \( p \), the graph is a small-world network (See Figure 1). This study uses the concept of “long-range edge” to explain the reason for small-world phenomenon of communication networks on microblogs in China.

![Figure 1. The transition from a regular network to a small-world network and finally to a random network using the random rewiring procedure (\( N=20 \) and \( K=4 \))](image)

Microblogs are among the most popular network platforms for information sharing, dissemination and acquisition based on user relationship, for they allow users to exchange small elements of content such as short sentences, individual images, or video links. Different versions of services and software with microblogging features have been developed. Among the most notable services are Twitter, Tumblr, FriendFeed, Plurk, Jaiku and identi.ca. Other leading social networking websites such as Facebook, MySpace, LinkedIn, Yahoo Pulse, Google Buzz, Google+ and XING also have their own microblogging feature, which are better known as “status updates”. Because foreign microblogging services such as Twitter, Facebook, Plurk and Google+ are censored in China, internet users face a different situation and use Chinese weibo services such as Sina Weibo and Tencent Weibo, which are also the most popular two microblogs in China. This study selects Tencent Weibo to conduct an empirical study on the small-world phenomenon of microblog communication networks in China.

**Data**

The empirical data of this study are collected from Tencent Weibo (http://t.qq.com.cn) on February 10, 2017. Tencent Weibo is one of the most popular microblogs in China whose slogan is “Your voice, echoes of the world”. Up to June, 2016, Tencent Weibo has approximately 540 million registered users and more than 100 million daily active users. On Tencent Weibo, several celebrity users have received more attention than a great many of
grassroots users, thus they take a more important role in the communication process of network information and public opinion. In our study, we select celebrity users with more than 6 million followers. We collect data using the snowball sampling method which is a kind of non-probability sampling method.

We randomly select a celebrity user who has more than 6 million followers, then look through his/her followed users, and record those users who have more than 6 million followers, thus we collect user information in the first round of sampling. We repeat the above process of sampling to each user of the first round, thus we collect user information in the second round of sampling. The celebrity user “Kaifu Li” (http://t.qq.com/kaifulee) is the starting point of this process of snowball sampling. We repeat the above process of sampling and finally collect information of 143 celebrity users in three round of sampling. The above 143 users come from Mainland China, Hong Kong and Taiwan and cover such broad areas as business circles, literature and art circles, sports circles and academic circles. It’s fair to say that we’ve got a representative sample of microblog users.

The “following” and “being followed” relationship between above 143 users can be signified in the form of 143×143 1-mode adjacency matrix $Z_{ij}$. The “following” relationship is a unidirectional one, thus the matrix rows signify the followers while the matrix columns signify the followed users, and the direction of edges derive from the followers to the followed users. For each one of above 143 users, if one user appears on another user’s list of followed users, then the value of corresponding matrix element is 1, otherwise the value is 0, and finally constructs a 143×143 binary adjacency matrix $Z_{ij}$. We present visually the adjacency matrix $Z_{ij}$ of this Microblog communication network using the drawing software NetDraw (see Figure 2).

Figure 2. The communication network of celebrity users on Tencent Weibo

The reason that user A follows user B can be summarized as follows: (1) user A is one of friends or acquaintances of user B in their actual social network, and the “following” relationship between user A and B is usually a bidirectional one in this case; (2) user B is a well-known figure in society whose words and deeds have extensive social influence; (3) user B is attracted by the worth and wit of what user A has to say, for example, the grassroots user “We Like to Tell Cold Jokes” has more than 18 million followers, and the “following”
relationship is usually a unidirectional one in the latter two cases. If user A follows user B, the direction of edge in their communication network is from A to B: A→B, while the direction of information dissemination is from B to A: B→A.

Methods

The information communication model of microblog is different from that of traditional media which has a linear propagation model (One to One), and it’s also different from that of network media which has a webbed propagation model (One to N), and it has a emerging circumnuclear propagation model (One to N to N) which allows information to disseminate in a geometrical progression in a short period of time. The information communicates through microblog depending on its users’ social networks which are sets of relationships between social actors. Each actor publishes information in his/her social network, so he/she is both sender and receiver of information, that is, each actor disseminates information through microblog depending on his/her own social network. Thus, we can conduct empirical study of information communication on microblogs with the theory and method of Complex Network Analysis (CNA) and Social Network Analysis (SNA).

Complex Network Analysis

The small-world network is defined as highly clustered like a regular network and yet with small average path length like a random network (Watts and Strogatz 1998). Therefore, the clustering coefficient $C$ and average path distance $L$ are the most important two parameters used for identifying a small-world network.

The clustering coefficient $C$ of a network

Suppose that node $i$ is connected to $k_i$ neighbors via $k_i$ edges. Then at most $k_i(k_i - 1)/2$ edges can exist between these $k_i$ nodes which occur when every neighbor of node $i$ is connected to every other neighbor of node $i$. The clustering coefficient $C_i$ of node $i$ is defined as the ratio of the actual number of edges $E_i$ between the $k_i$ nodes to the maximum possible number of edges $k_i(k_i - 1)/2$, that is, it can be expressed as:

$$C_i = \frac{2E_i}{k_i(k_i - 1)}$$

The clustering coefficient of the whole network is the average of the clustering coefficient of all nodes, that is, it can be expressed as:

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i$$

The average path distance $L$ of a network

The distance $d(i, j)$ between the nodes $i$ and $j$ in a network is defined as the number of edges in the shortest path connecting node $i$ and node $j$. The average path distance $L$ is defined as the number of edges in the shortest path between two nodes averaged over all pairs of nodes which can be calculated with the following formula:

$$L = \frac{2}{N(N+1)} \sum_{i \neq j} d(i, j)$$

where $N$ signifies the number of nodes within a network, $i$ and $j$ signify the nodes in a network, and $d(i, j)$ signifies the distance between the nodes $i$ and $j$. 
Social Network Analysis

The concept of “betweenness centrality” was firstly put forward by Freeman (1979) who applied it to analyze the importance of individuals in a social network. The betweenness centrality measures to what extent a node or an edge can control other nodes or edges. A node possibly performs an important mediation function if it exists amid several pairs of nodes. Even if the degree of this node might be relatively low, it is still a significant node within the network. Therefore, this study uses betweenness centrality as a measure to determine pivotal nodes and edges.

The node betweenness centrality

The node betweenness centrality measures to what extent a node can control the information communication between other nodes in a network. A node with a high betweenness centrality has the capacity to facilitate or limit interaction between the nodes it connects, and thus it has the capacity to control (bridge) the information dissemination within a network (Hung and Wang, 2010). For a given node \( n_i \), the node betweenness centrality \( C_B(n_i) \) of node \( n_i \) can be calculated as:

\[
C_B(n_i) = \sum_{j<k} \frac{g_{jk}(n_i)}{g_{jk}}
\]

where \( g_{jk} \) is the number of geodesics connecting nodes \( j \) and \( k \), while \( g_{jk}(n_i) \) is the number of geodesics connecting nodes \( j \) and \( k \) which pass along the node \( n_i \). The node betweenness centrality \( C_B(n_i) \) of node \( n_i \) usually can be normalized as:

\[
C'_B(n_i) = \frac{2C_B(n_i)}{(N-1)(N-2)}
\]

where \( C'_B(n_i) \) signifies the normalized node betweenness centrality of node \( n_i \), \( N \) signifies the number of nodes within a network, \( (N-1)(N-2)/2 \) is the maximum possible value of node betweenness centrality, which is that of the star graph consisting of \( N \) nodes.

The edge betweenness centrality

The edge betweenness centrality measures the number of an edge occurring on the geodesics between two nodes. It measures to what extent an edge can control the information dissemination in a network. For a given edge \( l_i \), the edge betweenness centrality \( C_B(l_i) \) of edge \( l_i \) can be calculated as:

\[
C_B(l_i) = \sum_{j<k} \frac{g_{jk}(l_i)}{g_{jk}}
\]

where \( g_{jk} \) is the number of geodesics connecting nodes \( j \) and \( k \), while \( g_{jk}(l_i) \) is the number of geodesics connecting nodes \( j \) and \( k \) which pass along the edge \( l_i \).

Results

The total number of edges in the Microblog communication network is 2,114, so its average number of edges per node equals 2,114/143=14.78. We construct a seldom network with almost the same average number of edges per node \( (K=15) \) using the social network analysis software Ucinet 6.2, then compute the clustering coefficient and average path length of the actual communication network and constructed seldom network. Table 1 shows the average path length \( L \) and clustering coefficient \( C \) for the actual network compared to the random network with almost the same number of nodes \( (N=143) \) and average number of edges per node \( (K=15) \). The average path length \( L \) of the actual network is almost the same with that of the random network, but the clustering coefficient \( C \) of the actual network is much more than that of the random network, which shows that the actual network possesses obvious small-
Table 1. Main parameter values of the actual network and random network

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Actual network</th>
<th>Random network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>143</td>
<td>143</td>
</tr>
<tr>
<td>Average number of edges per node</td>
<td>≈15</td>
<td>15</td>
</tr>
<tr>
<td>Clustering coefficient</td>
<td>0.419</td>
<td>0.190</td>
</tr>
<tr>
<td>Average path length</td>
<td>1.997</td>
<td>2.057</td>
</tr>
</tbody>
</table>

The pivotal nodes (users) which influence the information communication within the whole network can be determined by computing the node betweenness centrality of each node. Table 2 shows the top 10 pivotal users ranking with their node betweenness centrality whose node betweenness centrality are much more than most other users in the communication network (>650).

Table 2. The pivotal nodes (users) ranking with node betweenness centrality

<table>
<thead>
<tr>
<th>No. of nodes</th>
<th>Node betweenness centrality</th>
<th>Normalized node betweenness centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>2063.083</td>
<td>10.304</td>
</tr>
<tr>
<td>19</td>
<td>1650.207</td>
<td>8.242</td>
</tr>
<tr>
<td>34</td>
<td>1105.199</td>
<td>5.520</td>
</tr>
<tr>
<td>24</td>
<td>1062.935</td>
<td>5.309</td>
</tr>
<tr>
<td>11</td>
<td>1061.913</td>
<td>5.304</td>
</tr>
<tr>
<td>113</td>
<td>875.249</td>
<td>4.371</td>
</tr>
<tr>
<td>86</td>
<td>871.875</td>
<td>4.355</td>
</tr>
<tr>
<td>58</td>
<td>831.831</td>
<td>4.155</td>
</tr>
<tr>
<td>56</td>
<td>731.911</td>
<td>3.656</td>
</tr>
<tr>
<td>51</td>
<td>691.475</td>
<td>3.454</td>
</tr>
</tbody>
</table>

The pivotal edges (relations) which influence the information communication within the whole network can be determined by computing the edge betweenness centrality of each edge. Table 3 shows the top 12 pivotal relations ranking with their edge betweenness centrality whose edge betweenness centrality are much more than most other relations in the communication network (>142). It can be found that the majority of 12 pivotal edges originate from one of above 10 pivotal nodes, that is, one of two vertices of the 16 pivotal edges is likely to be the above 10 pivotal nodes.

Table 3. The pivotal edges (relations) ranking with edge betweenness centrality

<table>
<thead>
<tr>
<th>Edges</th>
<th>Edge betweenness centrality</th>
<th>Edges</th>
<th>Edge betweenness centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>54→23</td>
<td>282.733</td>
<td>23→116</td>
<td>146.896</td>
</tr>
<tr>
<td>113→141</td>
<td>260.100</td>
<td>51→125</td>
<td>146.266</td>
</tr>
<tr>
<td>108→58</td>
<td>154.279</td>
<td>121→115</td>
<td>145.857</td>
</tr>
<tr>
<td>23→51</td>
<td>152.542</td>
<td>23→114</td>
<td>144.764</td>
</tr>
<tr>
<td>23→113</td>
<td>150.245</td>
<td>13→11</td>
<td>143.339</td>
</tr>
<tr>
<td>86→113</td>
<td>149.738</td>
<td>23→37</td>
<td>143.153</td>
</tr>
</tbody>
</table>
Discussion
This study constructs a communication network of celebrity users on Tencent Weibo using the method of snowball sampling, compares the clustering coefficient \( C \) and average path distance \( L \) of the actual communication network with the corresponding random network which are the most important two parameters used for identifying a small-world network, and finds that the communication network presents the small-world property for its clustering coefficient \( C \) is much more than that of the random network keeping their average path distance \( L \) nearly equal. The pivotal nodes and edges of the communication network are found by analyzing the betweenness centrality of each node and edge, which play an important role in the process of information communication in the whole network. These pivotal edges usually originate from one of the pivotal nodes and they may act as the long-range edges in the communication network, which significantly reduce the average path distance at the same time maintaining a high clustering coefficient, thus the whole network present the small-world property. To determine the pivotal nodes and edges contributes to guide public opinion and to deal with emergencies through microblogging platform. The pivotal nodes frequently act as the “opinion leaders” in the process of information dissemination, with the aid of “long-range edges” of the communication network, internet public opinions can rapidly diffuse among the whole communication network on microblogs.

Acknowledgments
This work was supported by the National Social Science Foundation of China under Grants Nos.16BTQ060 and 15ATQ008, the Philosophy and Social Science Project of Heilongjiang Province in China under Grants No.15TQD003 and the University Basal Research Fund Project of Heilongjiang Province in China under Grants Nos. HDJ DY201615 and HDRC201614.

References
A bibliometric analysis of articles published by Chinese authors in library and information science journals of SSCI

Li Guojun\(^1\) and Qiu Xiaohua \(^2\)

\(^1\)ligj@lib.pku.edu.cn  
Peking University, Beijing (China)

\(^2\)cufeqiuxiaohua@163.com  
Central University of Finance and Economics, Beijing (China)

Abstract
This paper aims to explore the international development of research on library and information science in China, including the four regions of Mainland China, Hong Kong, Taiwan, and Macau. We conducted a bibliometric analysis of publications written by Chinese authors in library and information science based on Social Science Citation Index between 1993 and 2014. The result shows that China’s SSCI papers in LIS are growing at a rapid speed. After 2008, the Mainland China surpassed Taiwan to become the largest region by the number of publications in LIS. WuHan University, Taiwan University, City University of Hong Kong are respectively the most productive institutions in Mainland China, Taiwan, Hong Kong. Ronald Rousseau is the most cooperative scholar with Chinese scholars who has made great contribution to promote the internationalization of Chinese library and information science. There is still a big gap on citation among Mainland China and Taiwan, Hong Kong.

Keywords  Bibliometric analysis • China • Library and information science • Social science citation index

Conference Topic  
Mapping and visualization
The application of informetrics on evaluation

Introduction
With its fast economic growth, Chinese government put plenty of money into the education to promote the internationalization of education in recent years. R&D expenditure in China has risen steadily over the years, with an annual growth rate of 23 % during the 2000s (Wang 2015). Therefore, China has exhibited an exponential growth with its scientific research output. China is one of the countries with a large number of scientific papers which are indexed by the Science Citation Index-Expanded database. But China’s scientific research output varies greatly across disciplines. Compared with the natural sciences, the social sciences in China still have a certain gap on internationalization. So many colleges in China ask their teachers of social sciences to publish more papers on internationalizing journals, especially those indexed by the Social Science Citation Index database.

The library and information science in China has experienced a fast development in this background. More and more Chinese students go to foreign LIS institutes for further studies. The cooperation and communication between Chinese and foreign institutes in LIS is becoming more and more frequent. In order to explore the internationalization and the research status of LIS in China, a bibliometric analysis base on SSCI has been made. Many studies use bibliometric tools to reveal the development situation of scientific production in a specific domain. Huang Youliang et al(2015) conduct a retrospective bibliometric analysis of articles about traditional Chinese medicine research. Sui Xiaoyun et al (2015) performed a bibliometric analysis of published research papers related to the Mekong River during 1991-2012. The study also made the comparative analysis of publications among three regions from China to understand the development of LIS in China, including Mainland China, Taiwan, Hong Kong and Macau.
Methods

Data source
The data of this study is from the online database of the Social Science Citation Index, SSCI, among Web of Science Core Collection. Articles from the four regions of China were identified by authors’ addresses. The search phrase used to search address for this study was “China or Hong Kong or Taiwan or Macau”. Then, web of science category was used to refine the research results focusing on the category of information science & library science. Publications that document types are Article and Review were selected for this study.

Time span
The time duration of most bibliometric articles is five, ten, fifteen or twenty years. This method of selecting is easy to understand, but to some degree it is a bit subjective. In this study, we determined the time span of research according to number of articles each year. Prior to 1992, there are fewer SSCI indexed articles written by Chinese authors in LIS each year, the maximum number of documents is 11. After that, the number of articles written by Chinese authors is steadily increasing each year. So the time span of this study was identified from 1993 to 2014. In total 3160 research articles written by Chinese authors published from 1993 to 2014 in LIS were considered for this analysis with 103 articles excluded before 1992. The first author and corresponding author are considered to have more contribution to papers. Therefore articles with first author’s address from China were selected for this study. There were 66 articles without first author address information from Web of Science that were identified by corresponding author addresses. In last 2655 articles with first author’s or corresponding author’s addresses from China were selected.

Data analysis
The documents records of 3160 articles with full bibliographic information were downloaded in plain text format from web of science. We made a java program to deal with these .txt files. We converted these txt files to Excel files by the java program and then performed the bibliometric analysis.

Results

Total number of articles
From 1993 to 2014, Chinese authors published 2655 papers in SSCI journals of LIS. Table 1 shows article number of four regions of China. Among these papers, 1205(45.4%) were from Mainland China, 982(37%) from Taiwan, 464(17.5) from Hong Kong, only 4(0.1%) from Macau. The Mainland China has become a major contributor in the field of international library and information science in China.

Table 1 The total number of articles from Mainland China, Hong Kong, Macau, Taiwan 1993-2014

<table>
<thead>
<tr>
<th>Regions of China</th>
<th>Number of article</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainland China</td>
<td>1205</td>
<td>45.4%</td>
</tr>
<tr>
<td>Taiwan</td>
<td>982</td>
<td>37%</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>464</td>
<td>17.5%</td>
</tr>
<tr>
<td>Macau</td>
<td>4</td>
<td>0.1%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2655</strong></td>
<td></td>
</tr>
</tbody>
</table>
Distribution of publication yearly

According to SSCI there are 2655 articles written by Chinese authors during the period from 1993 to 2014 in LIS. Results of publication output by year are shown in Fig. 1. It is not difficult to find that the growth trends of publication in total and publication in Mainland China are similar. The 2002 is a dividing line. Before 2002, HK&MC is more than the number of documents in ML and TW. After that ML and TW increased over HK & MC, but in different way. Though from 2003 to 2008 TW is growing faster than ML, ML rises sharply after 2008 exceeding TW. On the whole, the number of publications in HK&MC is stable, and the growth is slow. TW maintains a steady growth rate after 2000. ML keeps a rapid growth rate after 2008 and the gap between ML and TW is gradually expanding since 2012.

Fig. 1 Output of articles published by authors from the Mainland China (ML), Hong Kong (HK) & Macau (MC), Taiwan (TW) by year during 1993-2014

Distribution of Journal

Through the analysis of the published journals, we can find out the preference of the Chinese authors in the SSCI journals of LIS. Table 2 to 4 list respectively the top 10 journals published by authors from Mainland China, Taiwan, Hong Kong and Macau. SCIENTOMETRICS is the largest number of published literature journals in Mainland China and Taiwan. The largest number of literature journals in Hong Kong and Macau is INFORMATION & MANAGEMENT. This shows that the Mainland China and Taiwan scholars in LIS are more concerned about the field of scientometrics. Hong Kong and Macau scholars are interested in information management. There are four journals that appear in all the three lists, JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY, INFORMATION PROCESSING & MANAGEMENT, JOURNAL OF INFORMATION SCIENCE, INFORMATION & MANAGEMENT.

Table 2 Top 10 journals published by authors from Mainland China

<table>
<thead>
<tr>
<th>Rank</th>
<th>Journal</th>
<th>IF 2014</th>
<th>Number of articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SCIENTOMETRICS</td>
<td>2.183</td>
<td>269</td>
</tr>
<tr>
<td>2</td>
<td>INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE</td>
<td>1.655</td>
<td>141</td>
</tr>
<tr>
<td>3</td>
<td>JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY</td>
<td>1.846</td>
<td>78</td>
</tr>
<tr>
<td>4</td>
<td>INFORMATION PROCESSING &amp; MANAGEMENT</td>
<td>1.265</td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td>JOURNAL OF INFORMETRICS</td>
<td>2.412</td>
<td>49</td>
</tr>
</tbody>
</table>
Table 3 Top 10 journals published by authors from Taiwan

<table>
<thead>
<tr>
<th>Rank</th>
<th>Journal</th>
<th>IF 2014</th>
<th>Number of articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SCIENTOMETRICS</td>
<td>2.183</td>
<td>130</td>
</tr>
<tr>
<td>2</td>
<td>INFORMATION &amp; MANAGEMENT</td>
<td>1.865</td>
<td>112</td>
</tr>
<tr>
<td>3</td>
<td>ONLINE INFORMATION REVIEW</td>
<td>0.918</td>
<td>96</td>
</tr>
<tr>
<td>4</td>
<td>ELECTRONIC LIBRARY</td>
<td>0.535</td>
<td>76</td>
</tr>
<tr>
<td>5</td>
<td>JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY</td>
<td>1.846</td>
<td>73</td>
</tr>
<tr>
<td>6</td>
<td>INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT</td>
<td>1.550</td>
<td>64</td>
</tr>
<tr>
<td>7</td>
<td>INFORMATION PROCESSING &amp; MANAGEMENT</td>
<td>1.265</td>
<td>61</td>
</tr>
<tr>
<td>8</td>
<td>JOURNAL OF INFORMATION SCIENCE</td>
<td>1.158</td>
<td>58</td>
</tr>
<tr>
<td>9</td>
<td>TELECOMMUNICATIONS POLICY</td>
<td>1.411</td>
<td>25</td>
</tr>
<tr>
<td>10</td>
<td>GOVERNMENT INFORMATION QUARTERLY</td>
<td>2.321</td>
<td>19</td>
</tr>
</tbody>
</table>

Most productive institutions

Table 5 shows respectively the top 5 institutions ranked by the number of published literature from Mainland China, Taiwan, Hong Kong and Macau. Wuhan University has an absolute leadership position in LIS in Mainland China with 148 articles. Wuhan University is
the first institution in Mainland China to carry out the education of library and information science, which has a wide range of influence at home and abroad in the field of library and information science. Chinese Academy of Sciences with 98 articles is in the second place in Mainland China. The following is Zhejiang University with 64 articles, Peking University with 62 articles and Tsinghua University with 49 articles. In Taiwan, National Taiwan University is the most productive university with 126 articles, City University of Hong Kong with 126 articles lead the list in Hong Kong and Macau territory. From the three top lists, the mainland universities are in a dominant position, WuHan University is the most productive institution in China.

Table 5: The top 5 productive institutions

<table>
<thead>
<tr>
<th>Rank</th>
<th>Mainland China</th>
<th></th>
<th>Taiwan</th>
<th></th>
<th>Hong Kong and Macau</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>institution</td>
<td>Number of</td>
<td>institution</td>
<td>Number of</td>
<td>institution</td>
<td>Number of</td>
</tr>
<tr>
<td></td>
<td></td>
<td>publications</td>
<td></td>
<td>publications</td>
<td></td>
<td>publications</td>
</tr>
<tr>
<td>1</td>
<td>WuHan University</td>
<td>148</td>
<td>National Taiwan University</td>
<td>126</td>
<td>City University of Hong Kong</td>
<td>136</td>
</tr>
<tr>
<td>2</td>
<td>Chinese Academy of Sciences</td>
<td>98</td>
<td>National Chiao Tung University</td>
<td>59</td>
<td>Hong Kong Polytechnic University</td>
<td>89</td>
</tr>
<tr>
<td>3</td>
<td>Zhejiang University</td>
<td>64</td>
<td>National Central University</td>
<td>51</td>
<td>Chinese University of Hong Kong</td>
<td>71</td>
</tr>
<tr>
<td>4</td>
<td>Peking University</td>
<td>62</td>
<td>National Chengchi University</td>
<td>48</td>
<td>University of Hong Kong</td>
<td>67</td>
</tr>
<tr>
<td>5</td>
<td>Tsinghua University</td>
<td>49</td>
<td>National Cheng Kung University</td>
<td>45</td>
<td>Hong Kong University of Science and Technology</td>
<td>55</td>
</tr>
</tbody>
</table>

Highly-cited papers

Highly cited articles are worth analyzing because of the potential association between high citation counts and high quality research. Thought the analysis of highly cited articles we can learn the general information about the issues on which Chinese scholars are focusing and doing research. Table 6 to 8 respectively present Top 5 highly cited papers from Mainland China, Taiwan, Hong Kong and Macau. The most cited article is from Taiwan which published in SCIENTOMETRICS in 2004 with 517 times. Highly cited papers in China are concentrated on INFORMATION & MANAGEMENT, SCIENTOMETRICS, INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT, INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE. The total citations of this top five articles in Mainland China, Taiwan, Hong Kong & Macau are respectively 373, 1823 and 1131. This shows that the literatures from Taiwan have a high influence in the discipline of international LIS, followed by the Hong Kong. The citation frequency of literatures from Mainland China is low. Compared to Taiwan and Hong Kong, there is a large gap on citations of papers in Mainland China.
### Table 6 Top 5 highly cited papers from Mainland China

<table>
<thead>
<tr>
<th>Rank</th>
<th>Title</th>
<th>Authors</th>
<th>Journal</th>
<th>Year</th>
<th>Times Cited</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Understanding customer satisfaction and loyalty: An empirical study of mobile instant messages in China</td>
<td>Deng, ZH (Deng, Zhaohua); Lu, YB (Lu, Yaobin); Wei, KK (Wei, Kwok Kee); Zhang, JL (Zhang, Jinlong)</td>
<td>INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT</td>
<td>2010</td>
<td>92</td>
</tr>
<tr>
<td>2</td>
<td>Global stem cell research trend: Bibliometric analysis as a tool for mapping of trends from 1991 to 2006</td>
<td>Li, LL (Li, Ling-li); Ding, GH (Ding, Guohua); Feng, N (Feng, Nan); Wang, MH (Wang, Ming-Huang); Ho, YS (Ho, Yuh-Shan)</td>
<td>SCIENTOMETRICS</td>
<td>2009</td>
<td>74</td>
</tr>
<tr>
<td>3</td>
<td>Bringing PageRank to the citation analysis</td>
<td>Ma, N (Ma, Nan); Guan, JC (Guan, Jiancheng); Zhao, Y (Zhao, Yi)</td>
<td>INFORMATION PROCESSING &amp; MANAGEMENT</td>
<td>2008</td>
<td>72</td>
</tr>
<tr>
<td>4</td>
<td>Assessment of world aerosol research trends by bibliometric analysis</td>
<td>Xie, SD (Xie, Shaodong); Zhang, J (Zhang, Jing); Ho, YS (Ho, Yuh-Shan)</td>
<td>SCIENTOMETRICS</td>
<td>2008</td>
<td>69</td>
</tr>
<tr>
<td>5</td>
<td>Data mining of cellular automata’s transition rules</td>
<td>Li, X (Li, X); Yeh, AGO (Yeh, AGO)</td>
<td>INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE</td>
<td>2004</td>
<td>66</td>
</tr>
</tbody>
</table>


### Table 7 Top 5 highly cited papers from Taiwan

<table>
<thead>
<tr>
<th>Rank</th>
<th>Title</th>
<th>Authors</th>
<th>Journal</th>
<th>Year</th>
<th>Times Cited</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Citation review of Lagergren kinetic rate equation on adsorption reactions</td>
<td>Ho, YS (Ho, YS)</td>
<td>SCIENTOMETRICS</td>
<td>2004</td>
<td>517</td>
</tr>
<tr>
<td>2</td>
<td>What drives mobile commerce? An empirical evaluation of the revised technology acceptance model</td>
<td>Wu, JH (Wu, JH); Wang, SC (Wang, SC)</td>
<td>INFORMATION &amp; MANAGEMENT</td>
<td>2005</td>
<td>395</td>
</tr>
<tr>
<td>3</td>
<td>Why do people play on-line games? An extended TAM with social influences and flow experience</td>
<td>Hsu, CL (Hsu, CL); Lu, HP (Lu, HP)</td>
<td>INFORMATION &amp; MANAGEMENT</td>
<td>2004</td>
<td>386</td>
</tr>
<tr>
<td>4</td>
<td>Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation</td>
<td>Hsu, CL (Hsu, Chin-Lung); Lin, JCC (Lin, Judy Chuan-Chuan)</td>
<td>INFORMATION &amp; MANAGEMENT</td>
<td>2008</td>
<td>320</td>
</tr>
<tr>
<td>5</td>
<td>Towards an understanding of the behavioural intention to use a web site</td>
<td>Lin, JCC (Lin, JCC); Lu, HP (Lu, HP)</td>
<td>INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT</td>
<td>2000</td>
<td>205</td>
</tr>
</tbody>
</table>

Table 8: Top 5 highly cited papers from Hong Kong and Macau

<table>
<thead>
<tr>
<th>Rank</th>
<th>Title</th>
<th>Authors</th>
<th>Journal</th>
<th>Year</th>
<th>Times Cited</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success</td>
<td>Markus, ML (Markus, ML)</td>
<td>JOURNAL OF MANAGEMENT INFORMATION SYSTEMS</td>
<td>2001</td>
<td>278</td>
</tr>
<tr>
<td>2</td>
<td>Investigating healthcare professionals' decisions to accept telemedicine technology: an empirical test of competing theories</td>
<td>Chau, PYK (Chau, PYK); Hu, PJH (Hu, PJH)</td>
<td>INFORMATION &amp; MANAGEMENT</td>
<td>2002</td>
<td>217</td>
</tr>
<tr>
<td>3</td>
<td>Modelling sustainable urban development by the integration of constrained cellular automata and GIS</td>
<td>Li, X (Li, X); Yeh, AGO (Yeh, AGO)</td>
<td>INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE</td>
<td>2000</td>
<td>216</td>
</tr>
<tr>
<td>4</td>
<td>Factors affecting the adoption of open systems: An exploratory study</td>
<td>Chau, PYK (Chau, PYK); Tam, KY (Tam, KY)</td>
<td>MIS QUARTERLY</td>
<td>1997</td>
<td>212</td>
</tr>
<tr>
<td>5</td>
<td>A perception-based model for EDI adoption in small businesses using a technology-organization-environment framework</td>
<td>Kuan, KKY (Kuan, KKY); Chau, PYK (Chau, PYK)</td>
<td>INFORMATION &amp; MANAGEMENT</td>
<td>2001</td>
<td>208</td>
</tr>
</tbody>
</table>


Coauthorship network

Coauthorship network is a very important form of social network. There are many studies conducted by researchers to reveal the development status and structure of disciplines. Erjia Yan et al. (2010) constructed the mapping library and information science in China from 2002 to 2007 by a coauthorship network analysis. There are many tools to construct and visualize the coauthorship network, such as VOSviewer (van Eck 2010, Waltman 2010), CiteSpace (Chen, 2004, 2006; Chen et al 2010), The Science of Science (Sci2, Sci2 team 2009). In this paper, we use the Sci2 to build the coauthorship network of Chinese authors indexed by SSCI in LIS. Sci2 is a new knowledge mapping tool designed by Katy Börner team from Indiana University. Sci2 is a modular toolset specifically designed for the study of Science. The tool allows users to create various bibliometric networks and maps based on Web of Science.

There are 4407 nodes (authors) and 9331 edges (coauthorship) among these authors in this data set. The average degree is 4.2, which means that the average of 4.2 authors in an article. There are 460 weakly connected components and 150 isolated nodes. The coauthorship network is not weakly connected. The largest connected component consists of 3261 nodes. All vertices and edges in a graph will be some problems, such as the display is not clear, the tag overlap. We extracted top 200 nodes by number of authored works from the coauthorship network. The top 200 authors collaborative network visualization results are shown in fig 2. The author with the largest number of collaborators (64) is Ronald Rousseau, who is the president of the international society for scientometrics and informetrics. The following is Huang Mu-hsuan (55) from Taiwan university. In the field of international library and information science, there are
two active groups in China, one is for the leadership of Ronald Rousseau, the other is for the leadership of Huang Mu-hsuan. Ronald Rousseau is very well-known foreign scholar in LIS of China, who has worked with many Chinese scholars and institutions, is committed to promoting the research of LIS in China and strengthening the relationship between China and the world in the field of library and information science.

Conclusion

This paper studies the international development of research on library and information science in China from 1993 to 2014, including the four regions of Mainland China, Hong Kong, Taiwan, and Macau. We have witnessed the growth of the Chinese literatures indexed by SSCI in LIS, especially the Mainland China, which has surpassed the Taiwan and Hong Kong regions. The Mainland China is the leader in the study of library and information science on the number of published literatures. However, the mainland scholars’ papers are cited less frequently, the international influence is low in LIS, and there is a certain gap between Taiwan and Hong Kong scholars. Therefore, the Chinese mainland scholars should continue to improve the quality and impact of papers. It is necessary to strengthen cooperation and exchanges with academic institutions from other countries and regions, thereby enhancing the overall quality of mainland scholars and promoting the internationalization of library and information science in China.

Fig. 2 Coauthorship network of Chinese authors in LIS journals of SSCI
References


Impact and usage indicators for the assessment of research in scientific disciplines and journals

Pei-Shan Chi¹, Wolfgang Glänzel²

¹peishan.chi@kuleuven.be
ECOOM, KU Leuven, Louvain, Belgium

²wolfgang.glanzel@kuleuven.be
ECOOM and Dep MSI, KU Leuven, Louvain, Belgium

glanzw@iif.hu
Dept Science Policy and Scientometrics, Library of the Hungarian Academy of Sciences, Budapest, Hungary

Abstract
Proceeding from results of an earlier pilot study usage and citation impact data from the 2013 annual volume of the Web of Science Core Collection (WoS) were collected and used to build indicators. Besides basic indicators on citation impact, usage counts and international collaboration, we applied the method of Characteristic Scores and Scales (CSS) to analyse the distributions of citations and usage counts to further test the relation between the usage and citation impact. The results could confirm and extend the results of the previous study to a larger set of subjects and countries. In addition, a new journal metric, the Journals Usage Index, is proposed to supplement journal impact measures by a usage based index.

Conference Topic
Usage data; The application of informetrics on evaluation; Indicators

Introduction
In a recently published pilot study, the authors investigated the relation and distributions of citations and usage counts for three countries with similar publication output and five selected fields in the sciences and social sciences as a pars pro toto exercise (Chi & Glänzel, 2017). The Web of Science Core Collection (WoS) daily-updated usage counts were analysed to measure the level of interest in a specific item. In order to do so, typical citation indicators have been calculated for their counterparts based on usage counts and compared with indicators of citation impact.

So-called usage statistics are provided by Clarivate Analytics (formerly Thomson Reuters) along with its Web of Science Core collection. According to Pringle (2015, p.1):

“This count measures the level of interest in a specific item on the Web of Science platform. The count reflects the number of times the article has met a user’s information needs as demonstrated by clicking links to a full-length article at the publisher’s website (via direct link or OpenURL) or by saving the metadata for later use.”

These counts may be an interesting supplement to citation indicators at practical any level of aggregation since unlike altmetrics indicators, which are devised to measure the impact on the general public, usage focusses again on communication among scholars (Pringle, 2015). The results showed that citations and usage counts in WoS correlate significantly. The application of Characteristic Scores and Scales (CSS) to the samples proved the usefulness and robustness of the method also in the context of usage distributions. In the present study, we aim at deepening the results of the pilots study by extending the dataset to a complete annual volume of the WoS, broadening their scope to disciplinary and journal analysis to gain reliable and consolidated new insights of mote general validity. We also propose a new journal indicator,
the journal usage index, based on usage counts alongside with the traditional journal impact measures. Although all indicators are calculated for the complete dataset, we present, because of space limitation, only a selection of subject fields, disciplines, journals and countries. Further data will be presented in an extended version of this study and are available from the authors on request.

Data sources and data processing

All indicators built in this project and calculated for this study are based on bibliographic items extracted from the Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI) and the Arts & Humanities Citation Index (A&HCI) of Clarivate Analytics WoS database. Only “citable items”, i.e., documents of the type article, letter and review have been taken into account. For this study the volume year 2013 was used with an observation period of three years, i.e., 2013–2015 for citation and usage data. At this point we have to mention that the WoS used for this study are not downloaded from the online version of the database but taken from custom datasets provided by Clarivate Analytics. Documents have been assigned to the 16 major fields and 74 subfields according to the modified Leuven-Budapest classification system (see Glänzel, Thijs & Chi, 2016).

Methods and Results

Usage-count based macro-level measures

Proceeding from the scenario of our previous study (Chi & Glänzel, 2017), we have first selected five major fields for the presentation and analysis of data in this paper as well. Later on, in the course of the study we will zoom in on subfields (disciplines) and finally on the journal level. We will present the selected fields, disciplines and journals as examples being part of broader research areas that cover research in the life sciences, natural sciences, mathematics and social sciences. We have processed data on all major fields, disciplines and journals. The data and results of which will be made available in the supplementary online material of a later extended version of the paper according to the policies and guidelines of the journals where the study will be published. However, we start the exercise with the highest hierarchical level to gain insight into the peculiarities of scholarly communication in different broad but still quite heterogeneous areas of scientific research. In the second part, the disciplinary analysis will also take the subject-specific peculiarities into account to extend the analysis to more homogeneous subjects.

The five selected major fields representing natural sciences, life sciences and social sciences, were selected to analyse the relations between usage and citation impact, also taking the effect of international collaboration into account. The fields are Chemistry, Mathematics and Neuroscience & Behavior; furthermore, we selected the diversified but focused topics of non-internal medicine specialties (Clinical and Experimental Medicine II in the Leuven-Budapest subject classification scheme – cf. Chi et al., 2015) and economic, political & legal studies within the social sciences (Social Sciences II in our classification scheme). Here we stress again that we only present selections of fields and countries in this study.

Table 1 gives the basic indicators for four of the five fields (neurosciences will be used later, in the context of journal indicators) and a selection of 15 out of the most active countries in all fields combined and the world reference standard. We use the following basic indicators: number of papers in 2013, the mean citation rate in a 3-year citation window (MCR), the share of uncited papers in the same window (c(0)), the mean citation rate of internationally co-authored papers (MCRint) along with the corresponding indicators based on the usage counts,
i.e., mean usage count (MUC), the share of papers with zero usage in the 3-year observation window \(u(0)\) and the mean usage count of internationally co-authored papers (MUCint).

![Scatter plots of MCR vs. MUR of countries with at least 100 papers in four fields.](image1)

**Figure 1.** Scatter plots of MCR vs. MUR of countries with at least 100 papers in four fields.

![Scatter plots of c(0) vs. u(0) of countries with at least 100 papers in four fields.](image2)

**Figure 2.** Scatter plots of c(0) vs. u(0) of countries with at least 100 papers in four fields.

However, before we have a closer look at the data shown in Table 1, we will analyse the more general trends based on all countries with at least 100 paper each in the fields under study. Figure 1 and Figure 2 show the correlations between MCR and MUR and between c(0) and u(0).
u(0). The relation between MCR and MUR is more influenced by subject than the relation
between uncited or unused papers. In Figure 1, papers in chemistry show a trend to have high
usage when they are highly cited. This trend is not that strong in mathematics, compared to
other fields. The low citation and usage data in the field may be a reason of this weak
correlation. By contrast, lowly cited and used papers are highly correlated with each other in
all the fields as shown in Figure 2. It reveals that “uninteresting” papers would not arise
attention from the community either in its citing or using behaviour.
In the first place we observe a generally higher usage than citation impact in Social sciences II,
where the factor for “translating” impact to usage amounts to 7.6 followed by chemistry with
5.6. In mathematics we find a more moderate situation: one citation “corresponds” to 3.4 usages
and clinical medicine II forms the low-end here with 1.9. The usage range between 1.9 and 7.6
already points to subject-specific communication patterns of scientists in different fields.
Table 1. Comparison of citation and usage indicator of 15 selected countries in 2013 with 3-year
citation/usage window in four subject fields according to the Leuven-Budapest classification
scheme [Data sourced from Clarivate Analytics Web of Science Core Collection]
Country

Papers

AUS (A)
BEL (B)
BRA (R)
DNK (K)
FRA (F)
DEU (D)
IND (I)
IRN (N)
ISR (L)
CHN (C)
POL (P)
SGP (G)
ZAF (Z)
CHE (H)
USA (U)

6,035 8.51 8.1%
2,873 8.42 9.2%
5,651 4.18 19.6%
1,679 8.06 6.8%
12,520 7.35 10.3%
17,088 8.35 11.1%
19,192 5.49 17.9%
8,456 5.03 19.9%
1,300 8.49 10.9%
73,514 7.47 16.3%
6,071 4.09 21.2%
2,982 14.56 6.3%
1,533 5.13 18.2%
3,548 10.00 8.3%
43,056 9.78 10.3%

World

282,238

MCR

Chemistry
c(0) MCRint MUC

6.67 16.7%

Country Papers MCR

44.89
41.80
27.50
41.06
39.55
39.92
26.50
23.64
43.82
49.11
23.67
75.95
25.83
44.00
47.44

2.89
2.57
1.91
2.13
2.29
2.55
1.98
2.57
1.69
2.20
1.95
2.55
2.00
3.12
2.53

29.7%
30.2%
38.0%
31.0%
32.5%
31.0%
43.3%
37.4%
41.1%
40.4%
38.0%
27.9%
39.8%
26.0%
33.3%

World

2.12 38.4%

u(0) MUCint
9.2%
9.7%
13.1%
10.3%
9.8%
10.5%
10.9%
11.0%
10.5%
10.4%
16.6%
8.5%
14.1%
10.3%
9.4%

8.15 37.44 11.3%

Mathematics
c(0) MCRint MUC

AUS (A)
1,415
BEL (B)
850
BRA (R)
1,291
DNK (K)
339
FRA (F)
4,625
DEU (D) 4,008
IND (I)
1,816
IRN (N)
1,867
ISR (L)
879
CHN (C) 13,833
POL (P)
1,378
SGP (G)
466
ZAF (Z)
477
CHE (H)
839
USA (U) 13,802
63,587

9.23
8.81
5.74
8.40
7.65
8.92
7.34
6.61
9.54
10.53
5.68
14.78
6.37
9.96
10.27

12,085
3,922
8,539
3,239
10,831
17,513
5,626
2,546
2,444
13,723
2,328
1,345
1,419
5,422
83,323

42.78 247,768

5.70
7.74
3.51
7.52
6.02
6.06
2.86
2.66
5.36
4.45
4.43
5.23
5.72
7.15
5.77

16.7%
13.4%
28.9%
12.2%
23.3%
19.7%
35.2%
33.3%
20.7%
20.0%
24.2%
17.3%
16.0%
14.0%
17.3%

7.73
9.41
7.17
9.28
9.88
9.09
5.98
3.89
8.90
6.35
8.72
6.56
7.26
8.18
7.61

4.50 22.8%

6.61

u(0) MUCint Papers MCR

3.21 10.33 23.8%
2.60 10.50 20.6%
2.24 6.46 28.3%
2.43 9.08 23.3%
2.51 5.66 33.5%
2.90 6.26 31.5%
2.52 6.30 34.5%
3.68 7.39 33.7%
2.03 4.28 37.7%
3.08 9.77 17.5%
2.14 4.02 31.3%
2.72 12.99 15.9%
2.36 5.48 31.4%
3.12 8.00 26.6%
2.86 7.56 29.4%
2.68

48.32
43.08
33.88
43.53
40.42
42.86
34.80
26.95
45.65
60.60
32.07
78.87
28.34
45.68
50.57

Clinical and Experimental Medicine II
Papers MCR
c(0) MCRint MUC
u(0) MUCint

7.25 28.2%

10.79 2,850
10.55
838
5.89
393
9.54
617
6.05 1,768
6.43 3,110
6.61
361
7.96
103
4.79
488
11.40 2,264
4.38
215
13.59
450
7.20
389
8.17
906
8.42 17,118
7.62 46,296

206

2.70
2.95
1.52
3.33
2.47
2.86
1.98
2.96
2.49
2.71
1.71
3.20
1.83
3.19
2.98

11.56
12.58
7.24
10.59
8.50
8.98
5.89
7.20
8.97
10.77
9.62
9.54
9.60
9.99
9.06

14.3%
13.3%
22.1%
12.4%
18.7%
15.3%
25.9%
22.7%
14.5%
12.1%
20.2%
14.2%
19.7%
14.7%
15.0%

12.96
13.49
10.65
12.27
11.55
11.37
9.24
8.86
11.90
13.14
12.71
10.62
10.32
10.42
10.65

8.46 17.7%

10.85

Social Sciences II
c(0) MCRint MUC
29.7%
28.8%
52.2%
24.6%
34.4%
30.5%
41.3%
31.1%
33.6%
30.4%
51.6%
25.8%
43.2%
27.5%
29.3%

2.38 36.4%

3.53
3.28
2.78
4.11
3.14
3.64
2.41
4.24
3.33
3.12
3.44
3.62
2.83
3.53
3.66

19.90
23.79
16.07
22.87
18.63
19.81
19.71
17.42
18.94
26.99
12.35
31.26
14.16
21.79
20.12

u(0) MUCint
15.1%
11.9%
19.8%
12.6%
16.0%
14.5%
15.0%
16.5%
12.7%
13.6%
26.5%
11.3%
17.7%
13.7%
16.5%

25.65
25.48
23.35
27.06
22.03
24.08
22.23
18.92
19.80
29.92
19.05
35.44
21.68
24.62
26.04

3.27 18.03 18.2%

24.36


Internationally co-authored papers generally have higher citation and usage impact than all the papers. The large increases of citation and usage caused by international papers are specially obvious in the social sciences. Compared to usage data, citations have the larger differences between these two types of papers in all the four fields.

Inspired by the relational charts for the presentation of relative indicators (see Braun, Glänzel & Schubert, 1985; Schubert & Braun, 1986), we show the plot of the Relative Citation Rate (RCR, see Schubert & Braun, 1986) vs. Relative Usage Count (RUC). For simplicity we have just divided the countries’ indicator values by the corresponding field standard of the world total. The “relational charts”, where the neutral value of 1.0 is indicated by horizontal and vertical lines and the coincidence of relative usage and citation impact by a diagonal line, are illustrated in Figure 3 for the four selected fields and the 15 countries. In general, developed countries are located in the upper right quadrant whilst developing countries are presented in the lower left quadrant.

Besides the country specific features, such as the low values in both citation impact and usage for Brazil, South Africa, Iran and Poland contracted by high values of Switzerland, Denmark, Belgium, Australia and Singapore, we find quite interesting subject-specific details for some countries reflecting national preferences and capacities. Iran’s high impact and adequate usage in mathematics is striking, which is contrasted by low indicator values in this field. The same applies to the usage data of Germany and France. Denmark’s citation impact is not outstanding here, either. Instead, it has much better citation performance in clinical medicine and the social sciences. Singapore has remarkable performance in terms of both indicators in the most fields except for clinical fields.

![Figure 3](image-url)

**Figure 3.** Relative usage and citation indicators of 15 selected countries in 2013 with 3 year citation/usage window in four subject fields according to the Leuven-Budapest classification scheme. Chemistry (top left), Clinical and Experimental Medicine II (top right), Mathematics (bottom left) and Social Sciences II (bottom right) [Data sourced from Clarivate Analytics Web of Science Core Collection]
Usage CSS

Characteristic Scores and Scales (CSS) are obtained from a recursive procedure of iteratively truncating a sample at its mean value, recalculating the mean of the truncated sample and continuing this procedure until it is stopped, when the pre-set number of thresholds is reached or no new scores can be obtained. We give only a short mathematical explanation here, while a more detailed and formalised description can be found in previous studies by the authors (e.g., Glänzel, 2007).

Let \( X \) denote the random variable represented by the sample. Putting \( b_0 := 0 \) as the very first characteristic score, we then obtain the subsequent scores as \( b_k := E(X|X \geq b_{k-1}) \) for all non-negative integer values \( k = 1, 2, \ldots \). Usually we stop the procedure at \( k = 3 \). We just mention in passing that for a Lomax/Waring distribution underlying the sample we obtain a very interesting property, particularly

\[
b_k = E(X|X \geq b_{k-1}) = a \cdot b_{k-1} + b_1 = N^{\alpha}(a^{k-1}) \text{ with } a = \alpha/(\alpha-1); \ k = 1, 2, 3,
\]

where \( \alpha \) is the “tail” parameter of the distribution. The probability of being in the intervals \([b_{k-1}, b_k)\) between two adjoining scores amounts to \([a^{-\alpha}]^{k-1}(1-a^{-\alpha})\). These intervals define the “performance” classes from “poor” (Class 1) to “outstanding” (Class 4).

Turning theory to practice, we proceed from a suited bibliometric sample or dataset. Note that the sample needs not necessarily be a random one. This could, for instance, be a set of papers published in a given discipline and the variable \( X \) stands for citation rates, or usage counts. If the tail parameter is in the neighbourhood of the value 3, we have \((1-a^{-\alpha}) = 0.7037\) for \( \alpha = 3.0 \) and obtain a theoretical distribution of objects over classes 1 through 4 roughly following a 70%–21%–6.5%–2.5% rule. This rule of thumb could be empirically confirmed in the literature (Glänzel, 2007; Albarrán & Ruiz-Castillo, 2011; Chi & Glänzel, 2017; Glänzel, Thijs & Debackere, 2018).

Taking into account the different standards of citation impact and (assumably) also of usage in the major research fields, we break down data into the more homogeneous discipline level since the calculation of the CSS scores and classes requires at least this level of granularity (cf. Glänzel, Thijs & Debackere, 2014). In order to illustrate the subject dependence of characteristic scores and the relative insensitivity of performance classes defined on them, we show first the \( b_k \) values for the WoS 2013 volume and twelve of the 74 subfields according to the Leuven-Budapest scheme in a three-year observation window based on both citation impact and usage counts in Table 2. The selected subfields comprise biosciences, internal and non-internal medicine, chemistry, physics, engineering, and the social sciences and one each from mathematics, geo- and space sciences and neurosciences. The number of papers assigned to these disciplines is sufficiently large as it ranges between roughly 15,000 and almost 60,000.

If we compare the scores of citation impact and usage counts in the same discipline, we find the same typical patterns as pointed to in the previous section. Here the two extremes are particle & nuclear physics with a factor close to the value 1, while in business, economics, planning both kinds of communication are separated by almost one order of magnitude. The steps from \( b_k \) to \( b_{k+1} \) are otherwise, both between metrics and across disciplines, roughly proportional. By contrast, the distribution of citations over the four classes that are defined by the \( b_k \) scores are strikingly insensitive to both the metric and the underlying discipline. This kind of robustness is in line with the observations published in the above-mentioned studies by Glänzel (2007) and Glänzel, Thijs & Debackere (2014).
Table 3 shows the corresponding class distributions for the same selection of disciplines as above. The conclusion drawn from these empirical distributions is in line with observations made by Albarrán and Ruiz-Castillo (2011) and Glänzel, Thijs & Debackere (2014). According to their observations, the share of poorly cited papers (Class 1) amounts to roughly 70%, the share of those assigned to Class 2 to about 21% and the share of papers in the upper two classes is 6%–7% and 2%–3% of all papers, respectively. Translating this model to usage, we could speak about a distribution of papers with poor, fair, remarkable and outstanding usage with similar properties (cf. Table 3). The most uncoordinated shares of four classes between citation and usage appear in physics and engineering disciplines. Even though the social sciences have the largest difference between the scores of two indicators, the shares of their publications in four classes keep stable.

Table 2. Comparison of characteristic scores for 12 selected subfields in 2013 according to the Leuven-Budapest classification scheme [Data sourced from Clarivate Analytics Web of Science Core Collection]

<table>
<thead>
<tr>
<th>Subfield</th>
<th>Papers</th>
<th>Citation score</th>
<th>Usage score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$b_1$</td>
<td>$b_2$</td>
</tr>
<tr>
<td>Analytical, Inorganic &amp; Nuclear Chemistry</td>
<td>41,253</td>
<td>5.54</td>
<td>12.59</td>
</tr>
<tr>
<td>Astronomy &amp; Astrophysics</td>
<td>17,494</td>
<td>9.06</td>
<td>21.96</td>
</tr>
<tr>
<td>Business, Economics, Planning</td>
<td>33,176</td>
<td>2.59</td>
<td>6.47</td>
</tr>
<tr>
<td>Cardiovascular &amp; Respiratory Medicine</td>
<td>37,815</td>
<td>6.27</td>
<td>16.21</td>
</tr>
<tr>
<td>Cell Biology</td>
<td>25,738</td>
<td>10.97</td>
<td>26.44</td>
</tr>
<tr>
<td>Electrical &amp; Electronic Engineering</td>
<td>56,829</td>
<td>3.42</td>
<td>8.93</td>
</tr>
<tr>
<td>Microbiology</td>
<td>54,401</td>
<td>6.63</td>
<td>15.19</td>
</tr>
<tr>
<td>Ophthalmology/Otolaryngology</td>
<td>15,148</td>
<td>3.61</td>
<td>8.67</td>
</tr>
<tr>
<td>Particle &amp; Nuclear Physics</td>
<td>15,125</td>
<td>6.61</td>
<td>16.06</td>
</tr>
<tr>
<td>Psychology &amp; Behavioral Sciences</td>
<td>38,528</td>
<td>4.59</td>
<td>10.64</td>
</tr>
<tr>
<td>Pure Mathematics</td>
<td>26,400</td>
<td>1.49</td>
<td>4.09</td>
</tr>
<tr>
<td>Sociology &amp; Anthropology</td>
<td>20,268</td>
<td>2.90</td>
<td>6.69</td>
</tr>
</tbody>
</table>

Table 3. Comparison of CSS classes for 12 selected subfields in 2013 according to the Leuven-Budapest classification scheme [Data sourced from Clarivate Analytics Web of Science Core Collection]

<table>
<thead>
<tr>
<th>Subfield</th>
<th>Citation class</th>
<th>Usage class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class 1</td>
<td>Class 2</td>
</tr>
<tr>
<td>Analytical, Inorganic &amp; Nuclear Chemistry</td>
<td>66.5%</td>
<td>22.6%</td>
</tr>
<tr>
<td>Astronomy &amp; Astrophysics</td>
<td>70.2%</td>
<td>20.5%</td>
</tr>
<tr>
<td>Business, Economics, Planning</td>
<td>67.4%</td>
<td>22.4%</td>
</tr>
<tr>
<td>Cardiovascular &amp; Respiratory Medicine</td>
<td>70.4%</td>
<td>21.4%</td>
</tr>
<tr>
<td>Cell Biology</td>
<td>69.2%</td>
<td>22.0%</td>
</tr>
<tr>
<td>Electrical &amp; Electronic Engineering</td>
<td>70.1%</td>
<td>20.2%</td>
</tr>
<tr>
<td>Microbiology</td>
<td>67.4%</td>
<td>23.4%</td>
</tr>
<tr>
<td>Ophthalmology/Otolaryngology</td>
<td>66.9%</td>
<td>22.5%</td>
</tr>
<tr>
<td>Particle &amp; Nuclear Physics</td>
<td>68.1%</td>
<td>22.4%</td>
</tr>
<tr>
<td>Psychology &amp; Behavioral Sciences</td>
<td>66.8%</td>
<td>22.7%</td>
</tr>
<tr>
<td>Pure Mathematics</td>
<td>69.1%</td>
<td>23.1%</td>
</tr>
<tr>
<td>Sociology &amp; Anthropology</td>
<td>63.4%</td>
<td>24.2%</td>
</tr>
</tbody>
</table>
The Journal Usage Index

At this point the question arises of whether appropriate journal indicators could be devised on the basis of usage counts. The first and most obvious journal metric that puts itself forward for defining a usage-count analogue, is, of course, the Impact Factors. Once we have a dataset based on one publication year but an observation window of three years, we define our usage indicator following this structure. In order to avoid confusion, we call the impact metric just Journal Impact Measure (JIM; cf. Glänzel & Moed, 2002). The analogous usage-based metric will be called Journal Usage Index (JUI).

Table 4. The ten journals ranking highest according to a 3-year journal impact measure (JIM) in 2013 and their Usage Index (UI). Top: Neurosciences, Centre: Information Science & Library Science, Bottom: Inorganic & Nuclear Chemistry [Data sourced from Clarivate Analytics Web of Science Core Collection]

<table>
<thead>
<tr>
<th>Neurosciences Journal</th>
<th>Papers</th>
<th>JIM</th>
<th>UI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nature Reviews Neuroscience</td>
<td>70</td>
<td>56.49</td>
<td>70.60</td>
</tr>
<tr>
<td>Nature Neuroscience</td>
<td>236</td>
<td>34.31</td>
<td>29.94</td>
</tr>
<tr>
<td>Neuron</td>
<td>363</td>
<td>31.12</td>
<td>22.58</td>
</tr>
<tr>
<td>Trends in Cognitive Sciences</td>
<td>65</td>
<td>30.68</td>
<td>50.02</td>
</tr>
<tr>
<td>Trends in Neurosciences</td>
<td>73</td>
<td>26.38</td>
<td>35.82</td>
</tr>
<tr>
<td>Journal of Pineal Research</td>
<td>67</td>
<td>22.69</td>
<td>25.85</td>
</tr>
<tr>
<td>Progress in Neurobiology</td>
<td>56</td>
<td>21.54</td>
<td>40.13</td>
</tr>
<tr>
<td>Acta Neuropathologica</td>
<td>136</td>
<td>20.07</td>
<td>16.20</td>
</tr>
<tr>
<td>Biological Psychiatry</td>
<td>270</td>
<td>18.82</td>
<td>22.77</td>
</tr>
<tr>
<td>Molecular Psychiatry</td>
<td>173</td>
<td>18.30</td>
<td>21.80</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LIS Journal</th>
<th>Papers</th>
<th>JIM</th>
<th>UI</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAMIA</td>
<td>221</td>
<td>7.31</td>
<td>17.08</td>
</tr>
<tr>
<td>Information Systems Research</td>
<td>59</td>
<td>6.08</td>
<td>65.41</td>
</tr>
<tr>
<td>Journal of Informetrics</td>
<td>102</td>
<td>4.94</td>
<td>30.74</td>
</tr>
<tr>
<td>IJIM</td>
<td>96</td>
<td>3.95</td>
<td>44.52</td>
</tr>
<tr>
<td>Scientometrics</td>
<td>259</td>
<td>3.79</td>
<td>54.69</td>
</tr>
<tr>
<td>Journal of Health Communication</td>
<td>96</td>
<td>3.78</td>
<td>25.85</td>
</tr>
<tr>
<td>Information &amp; Management</td>
<td>69</td>
<td>3.64</td>
<td>43.59</td>
</tr>
<tr>
<td>JASIST</td>
<td>164</td>
<td>2.98</td>
<td>31.12</td>
</tr>
<tr>
<td>Telematics and Informatics</td>
<td>112</td>
<td>2.90</td>
<td>22.13</td>
</tr>
<tr>
<td>IJGIS</td>
<td>138</td>
<td>2.88</td>
<td>23.70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inorganic Chemistry Journal</th>
<th>Papers</th>
<th>JIM</th>
<th>UI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordination Chemistry Reviews</td>
<td>175</td>
<td>27.50</td>
<td>126.77</td>
</tr>
<tr>
<td>Inorganic Chemistry</td>
<td>1,647</td>
<td>10.13</td>
<td>44.96</td>
</tr>
<tr>
<td>Organometallics</td>
<td>845</td>
<td>9.24</td>
<td>28.66</td>
</tr>
<tr>
<td>Dalton Transactions</td>
<td>1,936</td>
<td>8.64</td>
<td>38.64</td>
</tr>
<tr>
<td>Journal of Inorganic Biochemistry</td>
<td>205</td>
<td>6.14</td>
<td>34.86</td>
</tr>
<tr>
<td>European Journal of Inorganic Chemistry</td>
<td>590</td>
<td>5.30</td>
<td>40.02</td>
</tr>
<tr>
<td>Applied Organometallic Chemistry</td>
<td>126</td>
<td>5.22</td>
<td>26.33</td>
</tr>
<tr>
<td>Journal of Biological Inorganic Chemistry</td>
<td>91</td>
<td>4.92</td>
<td>28.70</td>
</tr>
<tr>
<td>Structure and Bonding</td>
<td>55</td>
<td>4.51</td>
<td>4.11</td>
</tr>
<tr>
<td>Journal of Fluorine Chemistry</td>
<td>216</td>
<td>4.50</td>
<td>18.71</td>
</tr>
</tbody>
</table>

210
The most obvious question is, of course, aimed at the correlatedness of the two measures. Intuitively, one would expect some but not too strong correlations since we have observed a different extent of correspondence between impact and usage already at the subject level. In order to analyse the correlation, we have first selected all journals assigned to three selected WoS subject categories, particularly, Neurosciences, Information Science & Library Science (LIS) and Inorganic & Nuclear Chemistry. As we have mentioned in the beginning of this section, we present the selection as examples and will make available data and results on more science fields, disciplines and journals in an extended version of the study.

Spearman’s rank correlation amounted to 0.584 for the neurosciences, 0.766 for LIS, and 0.732 for the chemistry discipline. This seems to be reasonable, but does not indicate a really strong correlation.

Table 4 shows the lists of top ten JIM journals in the three selected subject categories. In information science and inorganic chemistry the UIs of those journals are much more boosted from their JIMs than those in neurosciences. This finding reflects the subject-specific communication patterns of scientists in different fields mentioned in the usage-count based measures.

Finally, to deepen this results we decided to study another type of relationship (cf. Glänzel, 2010), namely by directly analysing the plot of the conditional expectation $E(\xi(t) | \zeta(t))$ against $\zeta(t)$, where $\xi$ represents the usage counts and $\zeta$ stands for the citation rates. In this context we mention, that this method has already successfully been applied to the correlation between citation impact and download statistics (Glänzel & Heeffer, 2014). Theoretically there are two extreme cases: The condition $E(\xi(t) | \zeta(t)) = const$ with $r = 0$ means uncorrelatedness. The other extreme is the identity, i.e., $\xi(t) = \zeta(t)$, which results in the trivial solution $E(\xi(t) | \zeta(t)) = \zeta(t)$. However, because of the properties of conditional expectations there exists always a function $f$, such that $E(\xi(t) | \zeta(t)) = f(\zeta(t))$.

![Figure 4. The plot of usage index vs. citation impact based on conditional means for journals with at most 100 citable publications for all fields combined (2013 with 3-year observation window) [Data sourced from Clarivate Analytics Web of Science Core Collection]](image)

We have applied this method to all journals (i.e., in all WoS categories) with at least 100 citable publications. This approach avoids the effect that many small journals affecting the regression and thus distorting the results. Figure 4 shows the plot of usage against impact, where journals
are now grouped if there impact values fall in the same intervals. Correlation is moderate and, on an average, two usages globally correspond to one citation.

Conclusions
We found a significant moderate correlation between citations and usage counts in WoS. The slope reveals interesting subject-specific details on the relationship between impact and usage: While in several fields impact and usage are of the same order of magnitude (e.g., physics disciplines, subfields of biosciences, mathematics and geosciences), usage exceeds citation impact in other disciplines by up to one order of magnitude (e.g., subfields of the social sciences and chemistry). The phenomenon of the social sciences does not strike us unexpectedly since citations in journal literature is less significant in these disciplines making here the application of bibliometrics problematic.

Furthermore, no causality should be assumed as there is no clear direction from usage to citations or vice versa. This is in line with earlier observations concerning the correlation between downloads and citations, no causality in one particular direction should be assumed (Glänzel & Heeffer, 2014). One should also keep in mind that downloads, usage and citations reflect different forms of communication: downloads and usage does not necessarily result in publishing, while citing journal literature and there is no evidence that citations induce interest and usage by colleagues. As mentioned by Wang et al. (2016), the “clicking and saving” WoS usage counts may not perfectly represent the usage data of “views and downloads” provided by publishers’ platforms. We will further study impact and usage patterns in relation to download statistics as of our pending future research task.

The application of CSS proved working in the context of usage counts as well as a robust profiling tools regarding the distributions over “performance classes”. It even keeps the robustness of usage and citation distributions in the fields which have large difference between two indicators, such as the social sciences. Basically, CSS properties are very similar in impact and usage keeping their subject-specific peculiarities and the “translation factor” between impact and usage we have already observed for the basic indicators.

Finally, the new indicator (the Journals Usage Index) may serve as an interesting supplement to journal citation measures. According to the expectation, there is a moderate correlation between the two type of metrics, where the disciplinary “translation factor” is kept but journal-specific patterns exhibit unexpectedly higher/lower usage counts resulting in some noticeable changes in the disciplinary ranking lists. The basic trend that could be observed was that the differences between two journal indicators are larger in the social sciences and basic sciences compared to life sciences.

Acknowledgement
The authors thank Clarivate Analytics for providing usage data of the 2013 annual volume of the Web of Science Core Collection.

References


Glänzel, W., Thijs, B. & Chi, P.S. (2016), The challenges to expand bibliometric studies from periodical literature to monographic literature with a new data source: The Book Citation Index. Scientometrics, 109(3), 2165–2179.


Appendix

Key to the country and subject abbreviations

<table>
<thead>
<tr>
<th>Country</th>
<th>ISO Code</th>
<th>Chart key</th>
<th>Field code</th>
<th>Sub-field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>AUS</td>
<td>A</td>
<td>B2</td>
<td>cell biology</td>
</tr>
<tr>
<td>Belgium</td>
<td>BEL</td>
<td>B</td>
<td>C1</td>
<td>analytical, inorganic &amp; nuclear chemistry</td>
</tr>
<tr>
<td>Brazil</td>
<td>BRA</td>
<td>R</td>
<td>E2</td>
<td>electrical &amp; electronic engineering</td>
</tr>
<tr>
<td>Denmark</td>
<td>DNK</td>
<td>K</td>
<td>G1</td>
<td>astronomy &amp; astrophysics</td>
</tr>
<tr>
<td>France</td>
<td>FRA</td>
<td>F</td>
<td>H2</td>
<td>pure mathematics</td>
</tr>
<tr>
<td>Germany</td>
<td>DEU</td>
<td>D</td>
<td>I1</td>
<td>cardiovascular &amp; respiratory medicine</td>
</tr>
<tr>
<td>India</td>
<td>IND</td>
<td>I</td>
<td>L1</td>
<td>business, economics, planning</td>
</tr>
<tr>
<td>Iran</td>
<td>IRN</td>
<td>N</td>
<td>M4</td>
<td>ophthalmology/otolaryngology</td>
</tr>
<tr>
<td>Israel</td>
<td>ISR</td>
<td>L</td>
<td>N2</td>
<td>psychology &amp; behavioral sciences</td>
</tr>
<tr>
<td>China PR</td>
<td>CHN</td>
<td>C</td>
<td>P5</td>
<td>particle &amp; nuclear physics</td>
</tr>
<tr>
<td>Poland</td>
<td>POL</td>
<td>P</td>
<td>Y2</td>
<td>sociology &amp; anthropology</td>
</tr>
<tr>
<td>Singapore</td>
<td>SGP</td>
<td>G</td>
<td>Z3</td>
<td>microbiology</td>
</tr>
<tr>
<td>South Africa</td>
<td>ZAF</td>
<td>Z</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Switzerland</td>
<td>CHE</td>
<td>H</td>
<td></td>
<td></td>
</tr>
<tr>
<td>USA</td>
<td>USA</td>
<td>U</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A New Approach to Evaluate the Impact of Papers with Combining Social and Academic Influence

Deng Shengli\textsuperscript{1} and Xiang Yang\textsuperscript{2}

\textsuperscript{1}victorydc@sina.com
Wuhan University, Wuhan (China)

\textsuperscript{2}daaayang@whu.edu.cn
Wuhan University, Wuhan (China)

Abstract
The evaluation of paper impact is one of the core contents in the informetric field, and citation analysis or Altmetrics are commonly employed to measure the influence of papers in academia. This study focuses on constructing a multi-dimensional evaluation system of paper impact with integrating academic and social influence of papers effectively. Such indicators as Relative Citation Ratio (RCR), H-index and Impact torque are used to evaluate correspondingly academic influence, while the social influence is assessed by Altmetrics. Finally, Fuzzy Analytic Hierarchy Process (FAHP) is employed to calculate the weights of each indicator. In current research, 50 highly cited papers in psychological field are selected as a case for verifying the introduced evaluating approach. The findings show that the novel evaluation method synthesizes the academic influence and social influence of the paper effectively. Comparing with traditional single indicator evaluation method, the new approach explains comprehensive impact of paper scientifically with integrating multi-dimensional indicators.

Conference Topic
The application of informetrics on evaluation

Introduction
In this highly competitive pursuit of scientific research achievement, evaluating the influence of papers is the most effective approach for assessing the contributions of researchers, which reflects the academic value of the papers and the degree of cognition of others. Traditionally, these judgments mainly depended on some defective indicators and peer review. With the rise of open access and the networking of scientific research communication promote the development of Altmetrics, which foster a new approach for the comprehensive evaluation of paper’s influence. At present, researchers have increasingly turned to numerical approaches and divided the evaluation of papers into two categories. One emphasizes the improved evaluation indicators based on the citation quantity of the paper, such as H-index (Schubert, 2009), Impact Torque (Wang & Ye, 2014), Mean Normalized Citation Score (Lundberg, 2007), Relative Citation Ratio (RCR) (Hutchins, Yuan, Anderson, & Santangelo, 2016) and Academic Trace (Ye & Leydesdorff, 2013), etc.; The other is to optimize Altmetrics, such as calculating on relativity of evaluation indicators and citations based on Altmetrics (Li, Thelwall, & Giustini, 2012), optimizing the evaluation system of influence (Bornmann, 2014b) and evaluating new types of academic achievements.

Through each of the quantitative methods mentioned above has strengths, their accompanying weaknesses are to measure the academic and social influence of papers separately. To fill in this gap, we construct a multi-dimensional evaluation system of paper impact with combining academic influence and social influence effectively, to adopt H-index, RCR and Impact Torque as the indicators of academic influence, meanwhile use partial indicators of Altmetrics as the indicator of social influence. With this purpose, the following two research questions are proposed:
(1) How to assess the academic influence and social influence of the paper, and how to combine them effectively?

(2) Is there a relationship between the academic influence and social influence of the paper? What is the relative importance between the H-index, RCR, Impact Torque and Altmetrics?

Related Work
Many alternative methods for evaluating the influence of paper has been proposed mainly based on the bibliometrics (Garfield & Merton, 1979; Hutchins et al., 2016; Leydesdorff & Opthof, 2011; Schubert, 2009). However, with the change of the mode of academic publishing and academic exchange, the evaluation indicators are becoming increasingly inefficient for on-side insight and the time lag. It cannot reflect the influence of papers objectively and comprehensively (MacRoberts & MacRoberts, 2013). In the new research environment, the rise of Altmetrics enriches the evaluation system, and divides the evaluation of the paper into two categories: based on citation analysis or Altmetrics.

Measuring the influence of papers based on citation analysis
Since Garfield firstly proposed to use the impact factor to evaluate the influence of journals, citation analysis has increasingly become a scientific field of study (Garfield & Merton, 1979). Moed further proposed a Relative Citation Rate used the average citation counts to all publications in the journals to quantify the impact of the paper group (Moed, Burger, Frankfort, & Van Raan, 1985). Hutchins improved this method and rename RCR to quantify the influence of a research article by making novel use of its co-citation network to field-normalize the number of citations it has received (Hutchins et al., 2016). Van Raan extended the indicators which used the field-specific normalization to replace the journal-specific normalization, and presented the Citations Per Publication/Field Citation Score (CPP/FCSm) indicators (Van Raan, 2006). Opthof and Leydesdorff argued this procedure only generates an effective indicator in the case of normal distributions, when giving the skewed distributions, one should average the observed versus expected values which are to be divided firstly for each publication (Opthof & Leydesdorff, 2010). And Egghe proposed to employ averages of ratios to replace ratios of averages for assessing the MNCS evaluation indicator (Egghe, 2012). Without changing the acronym of “MNCS,” one could define the “Median Normalized Citation Score,” this would relate the new crown indicator directly to the percentile approach (Leydesdorff & Opthof, 2011). On the other hand, Schubert used h papers which have been cited h times in cited literature as the H-indicator of the paper in order to reflect the direct and indirect influence of the papers (Schubert, 2009). Wang has recently employed a new indicator Impact Torque (M) to evaluate the impact of the paper on scholars, which is calculated by multiplying the number of authors who are affected by papers as the force A and the paper’s influence span time D (Wang & Ye, 2014).

Measuring the influence of papers based on Altmetrics
Although some of methods based on citation analysis have dramatically improved our theoretical understanding of papers’ evaluation, none have measured the social influence of the paper. Altmetrics offer new ways to measure the relevance of public and research output, which can be considered an interesting option for assessing the societal impact of papers (Piwowar, 2013). Its advantages and disadvantages have been explained by Bornmann, and he also discussed the definition of Altmetrics (Bornmann, 2014a). Furthermore, Bornmann (2014b) introduced Altmetric-based scientific evaluation system, which involves both academic influence and social influence. In order to validate the scientificalness of Altmetrics, Galligan and Dyas conducted an in-depth study on the data source and data
providing platform of Altmetric evaluation system (Galligan & Dyas-Correia, 2013). On the other hand, many comparative studies with citation analysis were proposed to demonstrate the usefulness of Altmetrics (Thelwall, Haustein, Lariviere, & Sugimoto, 2013). For instance, Zahedi randomly selected 20,000 articles in WOS and analysed the correlation between Mendeley's readership and the number of citations (Zahedi, Costas, & Wouters, 2014). Eysenbanch, Shuai, Pepe and Bollen explored the correlation between the number of citations and tweets, but there comes to no consistent conclusion from them due to the different data sources (Eysenbach, 2011; Shuai, Pepe, & Bollen, 2012). Finally, Ortega found a weak correlation between traditional bibliometric and Altmetrics indicators at the author level (Ortega, 2015).

In summary, the current evaluation system of the impact of papers mainly focuses on the development and improvement of single indicator, thus, the evaluation of the impact of papers is biased. Although some scholars have put forward the comprehensive evaluation of academic influence and social influence, they don’t construct an effective evaluation system to evaluate the comprehensive impact of papers. Based on the academic influence and social influence of paper, this study tends to construct the scientific evaluation system of paper’s influence by fuzzy analytic hierarchy process (FAHP) for uncovering comprehensive impact of papers with employing the RCR, H-index and Impact Torque to evaluate the academic influence, and utilizing Altmetrics to evaluate the social influence of paper.

Methodology and Data

Description of the Model

For evaluating the influence of papers, we should consider their academic influence and social influence synthetically (Rongyin.Z, Fengjiao.G, & Jie.T, 2016). But how to measure the academic influence and social influence effectively? Galligan pointed out that Altmetrics can better reflect the social value of papers and prove its relevance to the public (Galligan & Dyas-Correia, 2013). In terms of measuring academic influence, the citation counts of paper have been a direct indicator of the academic influence in previous studies. However, there is a time lag of publications, and RCR is a more effective citation measurement indicator which incorporates a customizable benchmarking feature that relates field- and time- normalized citations (Hutchins et al., 2016). Simultaneously, Schubert pointed out that the number of citations reflects the direct influence of papers, but ignores the indirect influence. So he came up with the H-index to measure the direct influence and indirect influence of papers (Schubert, 2009). Wang stated that the academic influence of papers is firstly reflected by the influence of the relevant scholars, and not limited to the citation counts (Wang & Ye, 2014).

After reviewing previous approaches, this study defines the influence of papers with the combination of academic and social influence, and selects RCR, Impact Torque and H-index as key indicators, which measure the direct influence, indirect influence and influence on the relevant scholars respectively. Social influence is measured by Altmetrics, such as Mendeley, Blog, CiteULike and other social indicators. A hierarchical multiple model will be built to evaluate the influence of papers (Figure 1). The weight of the indicators is determined by the FAHP which combined with the power method. Finally, paper’s influence is obtained by using the multi-objective linear weighting.
Academia Influence Indicator

We report here a comprehensive measurement of the academia influence, which select RCR, Impact Torque and H-index as key indicators. RCR puts the co-citation network of papers as a subject area (Figure 2), which provides a dynamic view for studying field. The analysis process is consisted of five steps:

1. Calculate the average number of citations per year. The formula is presented as follows:
   \[ Acr = \frac{\text{Citations}(p)}{\text{Age}(p)} \]
   Where Citations (p) is the total number of essays cited, and Age (p) is the number of years since the publication;
2. Generate the co-citation network of papers;
3. According to the quotation rate of periodicals in the common network, it is estimated that the anticipated number of citations in the field of paper. And \( Fcr \) is defined as:
   \[ Fcr = \frac{\sum jcr_i}{N} \]
   Where there are N papers in the total network and \( jcr_i \) is the reference rate of journals;
4. According to the linear relationship between Acr and Fcr, we can generate the expected number of citations per year (Ecr) in the area, where the paper belongs to. And the ratio of Acr to Ecr is the RCR of the paper.

Another indicator is Impact Torque which examine the influence scale of the paper and the effect of aging. The impact Torque \( M \) is defined as the product of the influence of scholar force \( A \) and influence span \( D \), its formula is shown as follows:
Where \( t_j \) is the time limit that the paper has been experiencing; \( c_j \) is the quoted number of the jth year; \( C \) is the total number of cited; and \( N \) is the total number of years the article has been published.

RCR and Impact Torque are adopted to measure the direct influence, meanwhile we use H-index to quantify the indirect influence. The specific mathematical expression is presented as follows:

Set \( TC \) as the a sequence of descending order of number of cited times \((1, 2, \ldots, r, \ldots, z)\) in order to express the order of the paper; And \( TC_r \) as the total number of paper \( r \) cited, then the following sequence is available:

\[
TC = (TC_1, \ TC_2, \ldots, TC_r, \ldots, TC_z)
\]

And \( TC_1 \geq TC_2 \geq \cdots TC_r \geq \cdots TC_z \). Thus, the mathematical meaning of \( h \) is:

\[
H = \max \{r; r \leq TC_r\}
\]

**Social Influence Indicator**

Studies have shown that Altmetrics can measure the social influence of the paper effectively, it has 15 indicators for the evaluation of papers, including Twitter, Facebook, Redditers, News, Blog, Mendeley, CiteULike, F1000, Google +, Pinterest, Q&A, Weibo, Peer review, Wikipedia and Policy documents. While there are too many invalid indicators with zero value, such as Pinterest, F1000, etc. In current research, we will remove those indicators, which have excessive null value, and choose CiteULike, Mendeley, Blog, Policy, Twitter, News, Facebook, Wikipedia as key factors to measure the social influence of the paper.

**Weight Calculation**

In current study, the weight of each indicator was determined by FAHP, which introduces fuzzy consistent matrix and does not need consistency test. At the same time, we use power method to calculate the ordering vector, reduce the number of iterations and increase the convergence rate to meet the requirements of the computational accuracy. Specific steps are shown as below:

- Constructing a finite relational matrix. Using the \([0.1 \sim 0.9]\) scale and establishing a priority judgment matrix \( F = (f_{ij})_{n \times n} \);
- Converting the precedence relation matrix to a fuzzy consistent matrix. \( R = (r_{ij})_{n \times n} \), \( r_{ij} = \frac{r_i - r_j}{2n} + 0.5 \), where \( r_i = \sum_{j=1}^{n} f_{ij}, i = 1, 2, \ldots n \).
- Calculating ordering vector by square root method;
- Using the power method to obtain a higher precision ordering vector. The complementary judgment matrix \( R = (r_{ij})_{n \times n} \) is transformed into a reciprocal judgment matrix \( E = (e_{ij})_{n \times n} \), where \( e_{ij} = r_{ij}/r_{jj} \). The sort vector \( W^{(0)} \) is taken as the initial vector \( V^{(0)} \) and iterated with the formula \( V^{(k+1)} = EY^{(k)}, V^{(k)} = \frac{V^{(k)}}{\|V^{(k)}\|_\infty}, k = 1, 2, \ldots \) If \( \|V^{(k+1)}\|_\infty - \|V^{(k)}\|_\infty < \varepsilon \), \( \varepsilon \) is the given error, then \( \|V^{(k+1)}\|_\infty \) is the maximum eigenvalue. The sort vector \( A \) is calculated as follows:

\[
A = \left[ \frac{V_{k+1,1}}{\sum_{i=1}^{n} V_{k+1,1}}, \ldots, \frac{V_{k+1,n}}{\sum_{i=1}^{n} V_{k+1,n}} \right]^T
\]

**Impact Calculation**
After obtaining the weights of each indicator by the FAHP, paper’s influence is calculated based on the multi-objective linear weighted calculation. The formula is shown as follows:

\[
\text{Impact} = \sum_{i=1}^{n} A_i \sum_{j=1}^{m} W_{ij} P_{ij}
\]

Where Impact is the evaluation of paper’s influence, \(A_i\) is the weight of the i-th indicator at criterion layer, \(W_{ij}\) is the weight of the j-th project hierarchy indicator under the i-th criterion, \(P_{ij}\) is the actual value of the jth project hierarchy under the i-th criterion.

Data collection and preprocessing

At present, the RCR is limited by paper PMID number, so current study chooses psychology papers as the analysis object. In addition, for calculating the paper Impact Torque, manual collecting data is need. Due to time limitation, only selected 50 psychology papers which are high cited are employed as sample.

As a case study, we select the Web of Science core set as the retrieval platform, psychology as the subject and the “ARTICLE” as document type, all publication and information about all citation report were downloaded and analysed. Sorting by the number of citations, top 500 literatures are selected as the preliminary sample and then exported to EndNote. After removing the papers without the DOI number, 440 papers are available. With retrieving the corresponding literatures in Altmetric.com according to the DOI, Altmetrics indicators and search results are obtained. The next step is to filter out the literatures without the PubMed, and 265 articles are left. Finally, sorting the literatures in descending order according to Alt Score, top 50 literatures are selected as a sample of our study. In addition, we retrieve the corresponding RCR value in the ICite website according to the PubMed number of literature. Finally, Z scores is used to standardize all the data.

Results

Weights

We invited one professor, two doctoral students and three master students in the field of information science of Wuhan University to formulate the initial priority relation matrix of the evaluation system, 3 reliable judgment matrices are obtained as follows:

The preferential relation matrix of the criterion layer: \(F_A = \begin{pmatrix} 0.5 & 0.7 \\ 0.3 & 0.5 \end{pmatrix} \)

The preferential relation matrix of the project hierarchy: \(F_{B_1} = \begin{pmatrix} 0.8 & 0.5 & 0.7 \\ 0.6 & 0.3 & 0.5 \end{pmatrix} \)

\[
F_{B_2} = \begin{pmatrix} 0.5 & 0.2 & 0.6 & 0.4 & 0.3 & 0.8 & 0.7 & 0.5 \\ 0.8 & 0.5 & 0.8 & 0.6 & 0.5 & 0.9 & 0.8 & 0.6 \\ 0.4 & 0.2 & 0.5 & 0.3 & 0.2 & 0.5 & 0.4 & 0.3 \\ 0.6 & 0.4 & 0.7 & 0.5 & 0.4 & 0.8 & 0.7 & 0.6 \\ 0.7 & 0.5 & 0.8 & 0.6 & 0.5 & 0.9 & 0.8 & 0.7 \\ 0.2 & 0.1 & 0.5 & 0.2 & 0.1 & 0.5 & 0.4 & 0.2 \\ 0.3 & 0.2 & 0.6 & 0.3 & 0.2 & 0.6 & 0.5 & 0.4 \\ 0.5 & 0.4 & 0.7 & 0.4 & 0.3 & 0.8 & 0.6 & 0.5 \end{pmatrix}
\]

In this paper, MATLAB is used to write FAHP process, the precision is set to 0.00001 and the maximum number of iterations is 1000 times. Then, the preferential relation matrix mentioned above is calculated as a parameter and input in program respectively to obtain the weight of indicators. These weights are calculated as follows:

\[
W_1 = W_{B_1}(W_{C_1}, W_{C_2}, W_{C_3})
\]

\[
= 0.6(0.2453, 0.4536, 0.3011),
\]
\[ W_2 = \mathcal{W}_{B_2}(W_{C_4}, W_{C_5}, W_{C_6}, W_{C_7}, W_{C_8}, W_{C_9}, W_{C_{10}}, W_{C_{11}}) = 0.4(0.1197, 0.1765, 0.0879, 0.1432, 0.1765, 0.0750, 0.0951, 0.1260) \]

**Papers’ Influence**

The influence of each sample was calculated by multi-objective linear weighting. In order to compare the consistency of our method with RCR, Impact Torque and H-index, the top 10 papers are selected as samples. Table 1 provides descriptive statistics of these papers’ influence, including the value and ranking of each indicator.

**Table 1 Evaluation parameters of the influence of paper**

<table>
<thead>
<tr>
<th>Alt</th>
<th>H</th>
<th>RCR</th>
<th>M</th>
<th>Impact</th>
<th>Alt-r</th>
<th>H-r</th>
<th>RCR-r</th>
<th>M-r</th>
<th>Impact-r</th>
</tr>
</thead>
<tbody>
<tr>
<td>301</td>
<td>137</td>
<td>209.8</td>
<td>148149.913</td>
<td>0.956887</td>
<td>14</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>190</td>
<td>108</td>
<td>208.8</td>
<td>247048.602</td>
<td>0.91328</td>
<td>23</td>
<td>20</td>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>133</td>
<td>17</td>
<td>786</td>
<td>26955.5803</td>
<td>0.853903</td>
<td>45</td>
<td>50</td>
<td>1</td>
<td>44</td>
<td>3</td>
</tr>
<tr>
<td>137</td>
<td>191</td>
<td>155.2</td>
<td>157103.847</td>
<td>0.766203</td>
<td>42</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>163</td>
<td>216</td>
<td>92.28</td>
<td>99852.183</td>
<td>0.745017</td>
<td>33</td>
<td>1</td>
<td>17</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>463</td>
<td>191</td>
<td>121.6</td>
<td>109012.026</td>
<td>0.672139</td>
<td>7</td>
<td>3</td>
<td>13</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>326</td>
<td>124</td>
<td>83.04</td>
<td>112809.998</td>
<td>0.66971</td>
<td>11</td>
<td>13</td>
<td>20</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>639</td>
<td>72</td>
<td>52.42</td>
<td>33475.4647</td>
<td>0.643432</td>
<td>3</td>
<td>35</td>
<td>29</td>
<td>29</td>
<td>8</td>
</tr>
<tr>
<td>152</td>
<td>134</td>
<td>164.2</td>
<td>118737.518</td>
<td>0.570128</td>
<td>37</td>
<td>10</td>
<td>6</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>135</td>
<td>137</td>
<td>137.6</td>
<td>135941.152</td>
<td>0.556807</td>
<td>43</td>
<td>9</td>
<td>8</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

In this paper, the standardized values of the indicators were compared, and the figure 3 is arranged in descending order of Impact. We found that the ordering of the indicators of the sample not only shows a certain degree of relevance, such as the article of the high influence, its Impact Torque, RCR ranking also higher, but also shows a certain degree of independence. Therefore, we checked for significant relevance between each indicator through the Spearman coefficient of SPSS. Because that the social impact indicators are excessive, here we only selected Blog, Twitter and Mendeley for correlation analysis, the correlation coefficient is in Table 2

**Figure 3 The distribution of the indicators of papers’ influence**

With the correlation analysis of each indicator, we found that the influence of the dissertation has a strong correlation among RCR, M and H-index, and H-index is also associated with more indicators. However, we found less significant relevance within the academia influence indicator and social influence indicator of papers.
### Table 2 The correlation coefficient of indicators of theses’ influence

<table>
<thead>
<tr>
<th>Indicator</th>
<th>RCR</th>
<th>H</th>
<th>M</th>
<th>Impact</th>
<th>Blog</th>
<th>Twitter</th>
<th>Mendeley</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RCR</td>
<td>1</td>
<td>-0.288*</td>
<td>.117</td>
<td>.516**</td>
<td>-0.200</td>
<td>.096</td>
<td>-0.150</td>
</tr>
<tr>
<td></td>
<td>Sig (2-tailed)</td>
<td>.042</td>
<td>.417</td>
<td>.000</td>
<td>.164</td>
<td>.507</td>
<td>.300</td>
</tr>
<tr>
<td>H</td>
<td>-0.288*</td>
<td>1</td>
<td>.585**</td>
<td>.437**</td>
<td>.002</td>
<td>-.344*</td>
<td>.241</td>
</tr>
<tr>
<td></td>
<td>Sig (2-tailed)</td>
<td>.042</td>
<td>.000</td>
<td>.001</td>
<td>.990</td>
<td>.015</td>
<td>.091</td>
</tr>
<tr>
<td>M</td>
<td>.117</td>
<td>.585**</td>
<td>1</td>
<td>.701**</td>
<td>-.145</td>
<td>-.120</td>
<td>-.018</td>
</tr>
<tr>
<td></td>
<td>Sig (2-tailed)</td>
<td>.417</td>
<td>.000</td>
<td>.000</td>
<td>.317</td>
<td>.406</td>
<td>.900</td>
</tr>
<tr>
<td>Impact</td>
<td>.516**</td>
<td>.437**</td>
<td>.701**</td>
<td>1</td>
<td>.210</td>
<td>.235</td>
<td>.227</td>
</tr>
<tr>
<td></td>
<td>Sig (2-tailed)</td>
<td>.000</td>
<td>.001</td>
<td>.000</td>
<td>.143</td>
<td>.100</td>
<td>.112</td>
</tr>
<tr>
<td>Blog</td>
<td>-.200</td>
<td>.002</td>
<td>-.145</td>
<td>.210</td>
<td>1</td>
<td>.527**</td>
<td>.347*</td>
</tr>
<tr>
<td></td>
<td>Sig (2-tailed)</td>
<td>.164</td>
<td>.990</td>
<td>.317</td>
<td>.143</td>
<td>.000</td>
<td>.013</td>
</tr>
<tr>
<td>Twitter</td>
<td>.096</td>
<td>-.344*</td>
<td>-.120</td>
<td>.235</td>
<td>.527**</td>
<td>1</td>
<td>.067</td>
</tr>
<tr>
<td></td>
<td>Sig (2-tailed)</td>
<td>.507</td>
<td>.015</td>
<td>.406</td>
<td>.100</td>
<td>.000</td>
<td>.644</td>
</tr>
<tr>
<td>Mendeley</td>
<td>-.150</td>
<td>.241</td>
<td>-.018</td>
<td>.227</td>
<td>.347*</td>
<td>.067</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Sig (2-tailed)</td>
<td>.300</td>
<td>.091</td>
<td>.900</td>
<td>.112</td>
<td>.013</td>
<td>.644</td>
</tr>
</tbody>
</table>

* Correlation is significant at the 0.05 level (2-tailed).

#### Discussion

**Compare with other evaluation methods**

This paper measures the academic influence and social influence of papers comprehensively by constructing a multi-dimensional influence evaluation model of papers. Comparing the evaluation results of our method with the Altmetric, RCR, H-index and Impact Torque of the top 10 articles, the evaluation of consistency between Altmetric.com and our method is 0.2, H-index is 0.5, RCR is 0.6, and Impact Torque is 0.8. In the comparison of consistency of evaluation results, Altmetric.com has the lowest consistency with our method while the academic influence indicator has a higher consistency, which could be due to the selection of partial Altmetric.com evaluation indicators. Figure 3 presents the distribution of each indicator, showing a general consistency with our method, which demonstrate that the new impact evaluation method of this paper takes into account the influence dimension reflected by the indicators synthetically. In general, new approach is more comprehensive and scientific than other evaluation methods.

**Assess the importance of each indicator**

The weights of each indicator in the evaluation system are obtained through the processing of FAHP. The results show that the weight of academic influence (0.6) in the evaluation method is higher than that of social influence (0.4), which may be due to the fact that the essence of papers is the academic achievement of the researcher. In the indicators of academic influence, the weight distribution is “RCR (0.4536)>Impact Torque (0.3011)>H-index (0.2453)”, which RCR overcomes the difference between discipline and time with the highest weight. Impact Torque evaluates the impact of the paper on the researchers from another angle, and its weight is slightly higher than H-index. As for the indicators of social influence, Twitter (0.1765), CiteULike (0.1765) and Blog (0.1432) have higher weights than others. Twitter is one of the top ten Internet access sites, while Blog is the first blog service provider, so both of them have...
a high weight of social impact. On the other hand, CiteULike is a document management tool, also reflects the academic influence of the paper to a certain extent, and therefore also has a high weight.

The independence and relevance of each indicator

In this paper, we compared each indicator to determine whether there were correlations between them, as shown in Table 2. We found a weak significant correlation between RCR and H-index (0.288), and no correlation was found between RCR and Impact Torque (0.177), which mainly because they measure the influence of different directions of paper. However, we found a moderate significant correlation between H-index and Impact Torque(0.585), which shows that H-index synthetically considers the direct influence and indirect influence of the paper. On the other hand, the correlation coefficients between Blog and Twitter, Blog and Mendeley, Mendeley and Twitter are 0.527, 0.347, 0.347 respectively, indicating that there is moderate significant correlation between social influence indicators. As the academic influence and social influence of the paper measuring the influence of the two dimensions, there is no significant correlation between the social influence indicator and the academic influence indicator. Therefore, in the evaluation system of paper influence, each indicator has a certain degree of relevance and independence, and can be used to measure academic influence and social influence of paper perfectly.

Conclusion

Current study constructs the multidimensional evaluation model of paper’s influence with extracting the effective indicators of measuring both academic influence and social influence of papers, and comparing the important degree and correlation between the academic influence and the social influence indicators. The new approach provides a novel evaluation perspective, which is not limited for using single indicator to reflect the influence of a paper, but integrating meaningful indicators to achieve multi-dimensional and multi-method combination. There are also several limitations in this study. For instance, fewer samples are employed in this paper. The further research mainly lies in the improvement of the indicators in this paper, as well as the expansion of the core indicators in the evaluation system, so as to form a unified evaluation method of paper influence.

References


Scholars on Twitter: who and how many are they?

Rodrigo Costas1, 2, Jeroen van Honk1, Thomas Franssen1

rcostas@cwts.leidenuniv.nl; j.s.van.honk@cwts.leidenuniv.nl; t.p.franssen@cwts.leidenuniv.nl
1 Centre for Science and Technology Studies (CWTS) Leiden University, Leiden (the Netherlands)
2 Centre for Research on Evaluation, Science and Technology (CREST), Stellenbosch University, Stellenbosch (South Africa)

Abstract
In this paper we present a novel methodology for identifying scholars with a Twitter account. By combining bibliometric data from Web of Science and Twitter users identified by Altmetric.com we have obtained the largest set of individual scholars matched with Twitter users made so far. Our methodology consists of a combination of matching algorithms, considering different linguistic elements of both author names and Twitter names; followed by a rule-based scoring system that weights the common occurrence of several elements related with the names, individual elements and activities of both Twitter users and scholars matched. Our results indicate that about 2% of the overall population of scholars in the Web of Science is active on Twitter. By domain we find a strong presence of researchers from the Social Sciences and the Humanities. Natural Sciences is the domain with the lowest level of scholars on Twitter. Researchers on Twitter also tend to be younger than those that are not on Twitter. As this is a bibliometric-based approach, it is important to highlight the reliance of the method on the number of publications produced and tweeted by the scholars, thus the share of scholars on Twitter ranges between 1% and 5% depending on their level of productivity. Further research is suggested in order to improve and expand the methodology.

Conference Topic
Altmetrics; Studies on the level of individual scientists

Introduction
Social media have become increasingly important as a tool for scholarly communication (Sugimoto et al., 2016). Academic social networking sites such as Mendeley, Academia.edu and ResearchGate are used by scientists to connect with peers, share (pre)prints and track the visibility of their work. Twitter is a popular microblogging platform in which diverse communities, such as journalists, students and scientists among others, interact and exchange information. Earlier research has studied discourses on Twitter in comparison to news discourses (Mondragon et al., 2017) and the use of Twitter amongst different scientific disciplines (Holmberg & Thelwall, 2014). A major drawback of these previous studies is their small scale. The majority of research on Twitter use amongst academics deals with 500 accounts or less (but see Hadgu & Jäschke, 2014). In general, there is a lack of identification methods that can be used to identify scientists on Twitter on a larger scale. Earlier work has used a variety of (labour intensive) methods to identify scientists on Twitter. The most commonly used method is based on selecting Twitter accounts manually (Veletsianos, 2012; Lulic & Kovic, 2012; Holmberg & Thelwall, 2014; Haustein et al., 2014; Hwong et al., 2016), often starting from an established set of scientists through snowball sampling (identifying for instance all scientists that follow or are followed by the established set of scientists). A second method is self-identification through survey research (Rowlands et al., 2011; Van Noorden, 2014; Collins, Shiffman, Rock, 2016). In this method, that potentially can be used to collect a larger group of scientists active on Twitter, researchers send out surveys to a group of scientists asking them, amongst others to identify their social media accounts. However, such method is always limited by response rate as well as the need to identify contact information for scientists before being able to contact them. A third method
is by identifying scientists through (conference) hashtags or Twitter accounts (Ross et al., 2011; Hadgu & Jäschke, 2014; Veletsianos & Kimmons, 2016). Hadgu and Jäschke (2014) have been especially successful by combining accounts of various conferences to obtain a large sample of Twitter accounts affiliated with computer science conferences. A fourth method is the use of lists compiled by Twitter users (Sharma et al., 2012). This approach has resulted in the largest set of scientists on Twitter to date compiled by Ke, Ahn and Sugimoto (2016). This study collected Twitter accounts of lists where the list as well as the Twitter biography contained a scientist title (e.g. psychologist, economist, PhD, researcher, etc.). They were able to collect 45,867 Twitter accounts of (self-identified) scientists. A fifth strand of research on Twitter users originates in computer science analyses on the demography of Twitter. Such studies involve far less manual labor and often work with very large datasets extracted from the Twitter API. These studies use metadata of Twitter accounts (such as the Twitter biography and location) as well as tweets and the network of followers and following to identify geographical location, age, gender and occupation of users (Mislove et al., 2011; Sloan et al., 2013; Sloan et al., 2015). This approach is successful in obtaining information about the demographics of Twitter, but is not able to identify individual scientists. The inclusion in lists and the presence of scientists titles in the Twitter biography section, as developed by Ke, Ahn and Sugimoto (2016) is a viable strategy, however this approach favours more established scientists who are included in Twitter lists far more often than junior scientists.

In this paper we take a different approach by creating a match between two universes of individuals: the set of authors recorded in the Web of Science database and the Twitter accounts recorded by Altmetric.com. Altmetric.com has developed a database of all tweets (and their related metadata) that include a (link to) a DOI (mostly related to scientific journal articles), which gives us a population of 2.6 million Twitter accounts, many of them being very likely related to scholars. Our aim is to identify all those Twitter accounts that belong to individuals among the 22 million disambiguated authors recorded in the Web of Science with at least a publication after the year 2005. We think this approach is more viable than the approaches described above for a number of reasons. First, drawing on these two datasets makes it possible to match scholars (determined by their authorship of publications covered in the Web of Science) and Twitter handles in a systematic manner, involving far less manual labour. Second, when we identify the Twitter account of an author, we can immediately connect this to other data (such as institution, field of activity, publication profile, citation profile, collaborations, etc.) related to the scientific author. Third, while it relies partly on self-identification (the author has to be on Twitter using—to some extent—her own real name and has to have tweeted a paper recorded in Altmetric.com) it uses a variety of factors to determine whether an account can be ascribed to a particular scholar, thus going beyond the mere identification of the Twitter user as a scholar in the Twitter biographical section. Fourth, we aim at validating the linkages between scholars and Twitter accounts based on a gold standard of valid set of scholars with Twitter accounts, for this using the self-reported Twitter accounts of researchers recorded in the public ORCID registry (https://orcid.org/).

Methodology

Data sources

The two data sources used for this matching are the Web of Science (WoS) database, and the Altmetric.com database. We use the author-name disambiguation algorithm developed by Caron & Van Eck (2014) and we work with a set of 22,642,206 disambiguated author names. From the Altmetric.com database we extracted all distinct Twitter accounts that have tweeted
at least a paper recorded in Altmetric.com until April 2016 (i.e. 2,622,116 distinct Twitter handles).

Matching author-Twitter names

In order to match the two different data sets – WoS disambiguated authors and Twitter user accounts – we started with some normalisation and cleaning of the data in order to harmonize the characters from both sources. In WoS, Roman characters are always used. However, in Twitter, scholars can (and often do) also use other character sets. There is, as such, a bias toward this Roman character set inherent in the matching (although automatic transliteration of other character sets could be a future strand of research).

Moreover, accents and diacritics are also generally elided in the WoS database while not in Twitter. This has been compensated by applying specific matching strategies for specific accents. For instance, umlauts in German are often not merely removed, but rewritten with an added e, so that for instance ü becomes ue. In our matching we have accounted for both û, u and ue in these cases. We have attempted to apply such flexible matching to all frequent diacritics and spelling variations. Another frequent example is allowing hyphens to match whitespace. These are often used interchangeably in double last names, and so we consider them interchangeable.

The data from the Twitter side has issues of its own too. We use two data fields recorded in the Altmetric.com database: the handle (or username), and the “full name” as the users have entered it in their Twitter account. Twitter makes no distinction between first and last name. In fact, sometimes users do not actually enter their real name in this field at all. However, we are not interested in these cases as we aim at researcher that quite clearly disclose their identity in their Twitter accounts. Still, when they provide their actual name, it can come in various guises. They can provide any combination of first name, initials, last name, maiden name, title and sometimes even institution, and these can occur in almost any order. By looking at the source data, we have tried to catch as many of these as possible within what we might call “format templates”. So if we find a name that is formatted “X.X. Xxx” (where X is any alphabetical character) we assume that the first two characters are initials and the last three are the last name. If the name is formatted “ XX Xxx” it will be the same, yet if the second X is in lowercase (“Xx Xxx”) it will be considered a first name and last name. And so on. Another complication to this name field is that it is limited by Twitter to twenty characters. Of course, full names can easily exceed this, and it is hard to predict how users deal with this cut-off. In any case, it is sure to increase the number of incomplete names.

Matching of the Twitter account with the author name can also be done based on the name from the handle. Handles on Twitter have a maximum length of fifteen characters and can include only a combination of alphanumeric characters and underscores (_). Of course, the absence of a whitespace character as well as other characters that might appear in a name (like dots and dashes) makes this field less reliable than the full name field. On the other hand, the restrictions can also help. In the case of many Chinese and Japanese researchers, for instance, the full name field contains their name in their own character set, whereas in their Twitter username they have been forced to use the Romanized version. Like with the full name, we use “format templates” here. We here assumed underscores to function as dividers between first and last names, or initials and last name, and we also took into account the sequence of upper- and lowercase characters (so that the username “XxxYyy” can be divided into first name “Xxx” and last name “Yyy”).
Based on these extraction processes and attempts to make matching as flexible as possible, the two data sets are finally connected and pairs of authors and Twitter users have been created.

Scoring the matches

Based on the previous matching approach, a total database with 503,599,561 pairs of author names and Twitter accounts has been created. The following step is to determine which of those pairs are correct. In order to do so, a scoring method based on different rules has been developed in order to select the most likely correct pairs of authors and Twitter accounts. This rule approach is inspired by a similar approach for author-name disambiguation developed by Caron & van Eck (2014). Several data elements that can happen both in Twitter data and in WoS data have been collected in order to weight the pairs. Different rules have been applied in order to score the occurrence of the different elements, and a final score by summing the different rule-based scores has been provided for each author-Twitter account pair. Table 1 presents a summary of the rules and their scores.

Table 1. Summary of the criteria and scores for the different elements matched.

<table>
<thead>
<tr>
<th>Rules</th>
<th>Matching event</th>
<th>Criteria</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>General matching</td>
<td>This is the basic matching between authors and Twitter accounts. The minimum matching element required is surname and initial. Thus, by default, all matches have at least 1 score.</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>Full name of author and tweeter</td>
<td>Frequent full name (&gt;30 scholars in the whole database with the same name)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Medium frequent full name (between 5 and 30 scholars in the whole database with the same name)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Infrequent full name (less than 5 scholars in the database with the same name)</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>First name</td>
<td>Frequent first name (&gt;145 scholars in the whole database with the same first name)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Medium frequent first name (between 145 and 12 scholars in the whole database with the same first name)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Infrequent first name (less than 12 scholars in the database with the same first name)</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>First single name</td>
<td>Very frequent first single name (more than 31 scholars in the database with the same first single name)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Infrequent first single name (less than 31 scholars in the database with the same first single name)</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>First single name penalization</td>
<td>When the author has a first name in the papers but such name does not appear in the Twitter name(s) at all.</td>
<td>-2$^1$</td>
</tr>
<tr>
<td>5</td>
<td>E-mail URL (in the Twitter account and as obtained from the e-mail server URL of the author)</td>
<td>Very frequent author URL (&gt;187 scholars are linked to the same e-mail server URL)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Medium frequent author URL (between 187 and 18 scholars are linked to the same e-mail URL)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Infrequent author URL (less than 18 scholars are linked to the same e-mail URL)</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>Organization name</td>
<td>Very frequent organization name (&gt;403 scholars are linked to the same organization)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Medium frequent organization name (between 403 and 20 scholars are linked to the same organization)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Infrequent organization name (less than 20 scholars are linked to the same organization)</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>City</td>
<td>Very frequent city (&gt;5,515 scholars in the database are linked to an affiliation in the city)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Medium frequent city (between 5,515 and 210 scholars in the database are linked to an affiliation in the city)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Infrequent city (less than 210 scholars are linked in the database to</td>
<td>3</td>
</tr>
</tbody>
</table>

$^1$ We penalize when an author uses her first name in her papers but uses a different one (or none) in the Twitter name.
As presented in table 1, there are a total of 14 rules, one of them with a negative score. The rules can be organized in four main groups.

1) **Rules based on the matching of the names of the authors and the Twitter names** (rules 0 to 4). These rules are based on a relative clear identification of both the authors and the Twitter users, and their matching. Rule [0] refers the general matching of authors and Twitter information done in the previous step. Thus, all potential matches get, in principle, a score of 1, although later on this can be changed with the negative rule.

2) **Rules based on the matching of individual-related information of the authors and the Twitter users** (rules 5 to 8). These rules are based on the matching of different elements (mostly institutional and geographical) provided both by the authors in their papers (e.g. affiliations, countries, e-mails) and the Twitter users (URLs, Twitter name and Twitter handle, geographical information provided in the Twitter profiles, biographical description in Twitter, sometimes including their city, country or also academic institution). In these rules, different scores are established depending on the frequency of the different elements in WoS. Thus, the higher the number of scholars that are linked to the same entity in WoS, the lower the score that will be attribute to that match.

3) **Rules based on the matching of activity-related information** (rules from 9 to 12). Here the rules are based on the publications, fields and journals of the authors and the publications tweeted by the Twitter account. The rationale is that the more a Twitter account (sharing a minimum name similarity with a given author) has tweeted the same papers of the author (i.e. ‘self-tweeting’ activity), papers from the same micro-fields, and/or from the same journals of activity of the author, the higher the chances that the author and the Twitter account correspond to the same person. Additionally, if a Twitter account has been

---

2 Micro-topics are defined as the fields obtained in the publication-level classification developed by (Waltman & Van Eck, 2012)
mentioned (by another Twitter user) together with a paper of the author with which it has been matched, it is very likely that the Twitter account corresponds to the author, in a sort of co-tweeted event (e.g., somebody in Twitter has tweeted a link to a paper and the handle of the author in the same tweet\(^3\)).

These rules have the highest scores. Activity rules are expectedly more accurate in the identification and discrimination of the correct matches as they provide information on more specific patterns of the relationship between the scholar and the Twitter account.

There is also an additional rule [13] that assesses the commonness of the matching. If an author is only matched to one Twitter account, the matching is weighted more positively than when the author is matched to numerous different Twitter accounts.

Based on the previous rules, there is also an additional rule, labelled as the ‘preferred rule’. It is not a scoring rule but a selection one. After performing the scoring of pairs of matched based on rules 0-13, each author (when matched to several Twitter accounts) is only assigned to the Twitter account with which it has the highest score. In case of ties (i.e. when an author is matched to several accounts with the same score) the author is kept linked to all of them. This rule is expected to reduce the noise caused by matches of authors with very similar names and activity profiles, working in the same institute, field, etc., which may cause that they get high scores in their combination with different Twitter accounts. It is a rule to increase the precision of the matching procedure.

**Gold standard, validation of the matching and final selection of author–Twitter handle pairs**

In order to determine the threshold of the score for the final selection of the most adequate matches, we have performed precision-recall analysis using a database of self-reported linkages of authors with their Twitter accounts. This ‘gold standard’ is based on the public ORCID data from years 2015 and 2016 (Paglione et al, 2015; Haak et al, 2016).

From these databases we found a total of 768 of different individuals that are also authors in the database of disambiguated scholars and have reported a Twitter account in their public ORCID profile. We further selected those individuals with a Twitter account in Altmetric.com (i.e. researchers that have tweeted at least one paper) finding a total of 631 different individuals. This means that about 82% of the scholars with a Twitter account in the gold standard are recorded (at least once) in the Altmetric.com database, thus suggesting the potential relevance of this method in order to be able to identify scholars\(^4\). Of the 631 we further discarded 4 scholars as they had their last publication before the year 2005. A total of 627 individuals were finally considered in the analysis. Based on the pairs of author names and Twitter names and in the scoring of the matches as described in table 1 we estimated the precision/recall values for those researchers in the gold standard. The results of such analysis are presented in Table 2. Table 2 presents the overall results of the number of scholars that have been paired with a Twitter account with a given score value (from \(\geq 3\) to \(\geq 6\)), together with the precision/recall values based on the gold standard.

<table>
<thead>
<tr>
<th>Total score</th>
<th>Distinct scholars matched with a Twitter account</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\geq 6)</td>
<td>134604</td>
<td>97.0%</td>
<td>72.3%</td>
</tr>
</tbody>
</table>

\(^3\) See an example here: [https://Twitter.com/wmijnhardt/status/781245999545212930](https://Twitter.com/wmijnhardt/status/781245999545212930)

\(^4\) Somehow it supports the idea that scholars that are on Twitter tend, at some point, to tweet or re-tweet a scientific publication, thus entering into the realm of altmetric sources (Haustein, Bowman & Costas, 2016).
As it can be seen in Table 2, precision increases with higher scores, while recall decreases. Basically, with scores higher than 5 we reach values of precision higher than 95%, but values of recall below 80%. Regarding recall, with a score low as 2 we get a value of 92%. In the discussion of recall it is important to keep in mind that not all the author-Twitter account recorded in ORCID are valid matches, for examples, there are Twitter accounts reported that are institutional or collective (e.g. labs, teams, departments, etc.). Therefore, the values of recall reported here have to be regarded as relatively conservative.

In this paper, in order to make a selection with a relative balance between precision and recall, and considering that precision is something to be preferred over recall (it is reasonable to argue that is best to have correct matches over more but noisy ones) a minimum value of 4 has been chosen as a good compromise. As a result, more than 385,000 individual scholars have been linked to a Twitter account. This number of scholars found in Twitter is the largest so far compared to any of the previous studies. Even if we would have selected a minimum score of 6, we would have obtained a set of almost 135,000 scholars on Twitter (i.e. almost three times more than in the largest study of identification of scholars on Twitter so far, cf. Ke, Ahn & Sugimoto, 2016).

**Analysis of the presence of scientists on Twitter**

In this section a general analysis on the presence of scholars with a Twitter account is presented. For this analysis we have focused on distinguishing those scholars that have a Twitter account as identified with the methodology explained above in contrast to those without a Twitter account. Only individuals with an article, review and letter and in the period 1980-2015 and with at least a publication after 2005 have been considered in the analysis (a total of 17,332,510 disambiguated scholars). A total of 334,856 (2%) of them have been matched with a Twitter account with a score of 4 or higher.

In Figure 1 the share of researchers that have a Twitter account controlling by production is presented. As it can be seen the share of scholars on Twitter increases with the number of publications, ranging between 1% and 5%. It could be argued that this is a sign that more research active scholars are also more likely to be active on social media. However, considering that some of the scoring rules (i.e. activity rules – 9,10 and 12) are dependent on the number of publications of the authors, the outcomes of the analysis could also indicate that researchers with more output are more likely to be linked to their Twitter account in our methodology.

In Figure 2 the share of scholars with a Twitter account by scientific domain is presented. In order to classify scholars into major scientific domains the same approach as used by Larivière & Costas (2016) has been employed here. Thus, individuals are assigned to their main domain of activity based on their number of publications, and in case of ties (i.e. a scholar linked to several domains with exactly the same number of publications) a random assignation is applied. Results show how scholars with Twitter accounts are prominent in disciplines such as ‘Social and Behavioral Sciences’ (above 5%) as well as ‘Law, Arts and
Humanities’ (almost 4%). The shares in the ‘Medical and Life Sciences’ (about 2%) as well as in the ‘Natural Sciences’ (just above 1%) are much lower.

Figure 1. Overall number of scholars and percentage of them with a Twitter account, controlling by production

![Graph showing the overall number of scholars and percentage of them with a Twitter account, controlling by production.](image)

Figure 2. Percentage of scholars with a Twitter account by main domain of activity

![Bar chart showing the percentage of scholars with a Twitter account by main domain of activity.](image)

In Figure 3 an analysis on the age of the scholars is presented. In this case, age is determined by the year of first publication of the scholars. This year has been proven to be the best overall
proxy of the academic age of scholars (Costas, Nane, & Larivière, 2015). Controlling for the number of publications we find that scientists with a Twitter account tend to be on average younger than those without a Twitter account. Not surprisingly, scholars with a lower number of publications tend to be younger overall than those with more publications, although the pattern of younger scholars on Twitter is observed for all groups of productivity.

Figure 3. Average Year of first publication (YFP) of scholars with a Twitter account vs. those without a Twitter account, controlling by the number of publications

Conclusions and further research
In this paper we present a novel methodology for identifying scholars with a Twitter account. By combining bibliometric data from WoS and Twitter accounts from Altmetric.com we have obtained the largest matching of individual scholars with Twitter users made so far. Our methodology consists of a combination of matching algorithms, considering different linguistic elements of both author names and Twitter names; followed by a rule-based scoring system that weights the common occurrence of several elements related with the names, individual elements and activities of both Twitter accounts and scholars matched. The methodology developed presents interesting advantages with respect to previous approaches:
- It is a systematic approach that can be applied to large sets of researchers. It relies on the role of scholars are publishing authors and is not dependent on aspects such as the biographical descriptions of the Twitter users or their presence in lists (cf. Ke, Ahn & Sugimoto, 2016).
- It is subject of validation by external gold standard, in this case a gold standard based on data provided by ORCID has been employed, but the analysis of other golden sets would be also possible.
Once a Twitter account of an author is identified, it is possible to link this with additional bibliometric data (e.g. affiliations, scientific domain, citation impact, collaboration patterns, etc.) related to the scientific author. It is of course also possible to extract data from the Twitter handle, thus being able to incorporate information on the online activity of the scholar (e.g. followers, followees, (re)tweeting activity, hashtags, etc.). This opens a unique possibility to exhaustive studies on the activities that scholars are performing in social media as well as in their publications.

The main limitation of this approach is its reliance on WoS (see also Ke, Ahn & Sugimoto, 2016) and Altmetric.com, which means it can only be used to identify scientists that publish in journals included in WoS. This means we would fail to identify a larger amount of humanities and social science scholars active on Twitter but not active in WoS. Similarly, scholars who are on Twitter but not tweeting scientific outputs (or tweeting outputs not properly tracked by Altmetric.com) would also be excluded. Also, activities in social media platforms different than Twitter are not considered in this analysis (e.g. the Chinese Weibo).

In any case, results show positive values in terms of precision and recall suggesting a strong validity of this methodology. More than 385,000 scholars can be linked to a Twitter account. In general, our results indicate a presence scholars on Twitter between 1% and 5% in the overall population of scholars in the WoS. The numbers of scholars on Twitter vary by levels of productivity, thus researchers with higher levels of production have also a stronger presence on Twitter, this very likely also caused by the reliance of the method on the number of publications of scholars. By disciplines we find a strong presence of researchers from the Social Sciences and the Humanities. Natural Sciences is the scientific domain with the lowest level of scholars on Twitter. Researchers on Twitter tend to be younger than those that are not on Twitter. These results align with those reported by Ke, Ahn & Sugimoto (2016) who also reported a higher presence of Social Sciences and Historians on Twitter, and lower levels of Life and Natural Sciences as well as Mathematicians. The high level of Humanist scholars is remarkable, considering that previous studies found that the Humanities is not one of the strongest scientific domains in altmetrics (Costas, Zahedi, & Wouters, 2015) as well as the fact that publications of scholars in this domain are least likely to be covered in the WoS.

The approach presented here also opens new and diverse venues of further research:

- Improvement of the matching and scoring algorithms. An important element that will need to be explore in the future is how to improve the matching methodology. For example, how to better match names in other non-Roman alphabets, the better parsing of Twitter names, a better matching and scoring the pairs of authors and Twitter users, etc. All these are relevant research aspects that will need more attention in the future. Additionally, the creation and evaluation of additional gold standards (i.e. validated lists of authors and their Twitter and other social media accounts) will be a necessary step in order to validate future approaches of identification of scholars on social media.

- Incorporation of other databases. Including other bibliometric databases (e.g. Scopus, Google Scholar, Microsoft Academic) would increase the population of publications and scholars. Also, the inclusion of other altmetric data sources (e.g. PlumX Analytics) as well as working with social media platforms directly (as done in Ke, Ahn & Sugimoto, 2016) would increase the possibilities to identify scholars on social media platforms.

- Demography of scholars on Twitter. Based on the current research, further analysis should focus on more deeply studying the demography of scholars on Twitter. Thus, considering the combination of both bibliometric and Twitter data it will be possible to analyse the countries of the scholars on Twitter, their gender or their subdisciplines, among other perspectives (e.g. age, collaboration, mobility, etc.).
- Moreover, the combination of bibliometric and altmetric information opens a clear path to study of the relationship between bibliometric performance and Twitter and social media activity. Thus, questions related with the impact (citation and altmetric) of those scholars on Twitter compared to those who are not on Twitter will be key future questions.
- Similarly the analysis of the activities and interactions of scholars on Twitter will be an excellent method to better determine and contextualize the interactions that scholars are maintaining with other societal stakeholders, as suggested by Robinson-Garcia, van Leeuwen, & Rafols (2017).

It can be conclude that this approach represents as a qualitative first step towards a large identification of the science-Twitter universe in a systematic way combining both bibliometric and altmetric sources. A combination of our approach with the list approach (as in Ke, Ahn & Sugimoto, 2016) as well as an analysis of the biographies of the followers and followings of the identified scholars might help to counter the limitations of the current approach, opening the path to a more complete perspective of the true engagement of scholars on social media platforms.

Acknowledgements
This work has been supported by Eurostars-2 funded project SIA Graph. Rodrigo Costas was partially supported by funding from the DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP) (South Africa). The authors acknowledge the help by Josh Brown, Adéniké Deane-Pratt and Tom Deranville from ORCID in obtaining the gold standard database and the comments and feedback received from Cassidy Sugimoto and Vincent Lariviè re on early discussions of this paper.

References


Haak, Laurel; Brown, Josh; Buys, Matthew; Cardoso, Ana Patricia; Demain, Paula; Demeranville, Tom; Duine, Maaike; Harley, Stephanie; Hershberger, Sarah; Krznarich, Liz; Meadows, Alice; Miyairi, Nobuko; Montenegro, Angel; Paglione, Laura; Pessoa, Lilian; Peters, Robert; Monge, Fran Ramirez; Simpson, Will; Wilmers, Catalina; Wright, Douglas (2016). *ORCID Public Data File 2016*. figshare. https://doi.org/10.6084/m9.figshare.4134027.v1


Paglione, Laura; Peters, Robert; Wilmers, Catalina; Simpson, Will; Montenegro, Angel; Ramirez Monge, Fran; Tyagi, Shobhit; Krznarich, Elizabeth; Demeranville, Tom; Brown, Josh; Miyairi, Nobuko; Buys, Matthew; Cardoso, Ana; Sethate, Cheryl; Haak, Laurel (2015). ORCID Public Data File 2015. figshare. https://doi.org/10.6084/m9.figshare.1582705.v1


Abstract
This article studied the usability of altmetrics in the academic evaluation based on the data from Altmetric.com. This study was carried out in five steps. Firstly we sorted out all the related studies to make exhaustive understanding of the existing methods and conclusions for evaluating the usability of altmetrics. Secondly, the evaluation method was proposed and the dataset was constructed, which consisted of 18 altmetrics indicators of papers from 10 disciplines and published in recent 10 years. Thirdly, for every indicator we studied the distribution of coverage and mean count over time to analyze the indicators’ accumulation characteristics and focus on early literature. In the fourth part, the value of every indicator and the redundancy among values were studied. The approaches included correlation analysis between altmetrics indicators and citation, correlation analysis with each other among altmetrics indicators, and PCA among all indicators. In the last, we analyzed some factors that may affect the usability of altmetrics, such as discipline, JIF, country, language and so on.

The conclusions of this article are as follows. Some altmetrics indicators just pay attention to newly published papers, while other indicators could have wider coverage. This situation causes different data accumulation process. It’s common to find redundancies among indicators in value. The value reflected by altmetrics is various and not limited to academic value. Besides, the application of altmetrics in academic evaluation must consider discipline, JIF, language, country and so on.

Conference Topic
Altmetrics

Keywords Altmetrics, Social Media, Online Academic Communication, Academic Evaluation

Background
Peer review and citation analysis are the two main methods for academic evaluation. Peer review is a qualitative evaluation method which may be affected by the professors’ knowledge area, academic vision, or even academic morals (Cole, Cole & Rubin, 1977). As a quantitative evaluation method, citation analysis could avoid the problems above, but as it only relay on the citation relationship among papers, it has other shortages such as evaluation delay (Buschman & Michalek, 2013), measure limitation (MacRoberts & MacRoberts, 2010), and participations limitation for non-academic circles. Under the development of social media and open access, the academic communication form is constantly changing, with online academic communication becoming more and more important, and activities such as reads, download, use and collect should also be considered. So is the form of academic achievements. Besides papers, there are the blogs, presentations, software and so on. Peer review and citation analysis could no longer satisfy the needs of academic evaluation in new environment. Thus, many researches have tried to introduce some new indicators for the supplement of the existing evaluation system.

Altmetrics, short of alternative metrics, was put forward in 2010 by Jason Priem. It is the most well-known new form of metrics with great development in both theory and practice in the past several years. It has many advantages compared to traditional bibliometric (Piwowar & Priem, 2013). Indicators of different types are included to reflect actions like mention,
collection, recommendation and usage. The timeliness of many indicators is better than citations, while the forms of achievements they could measure are diverse. Nevertheless, altmetrics also have many shortcomings. Haustein (2016) summarized three main shortages as heterogeneity, data quality and dependencies. The data from websites makes it hard for altmetrics indicators to be as rigorous as citation. Many attentions achievements got from social media are the reflections of popularity or visibility on the Internet without deep understanding of their content. There are still many achievements in lack of uniform identify, which influence the accuracy and coverage of indicator data. All these problems are obstacles for the spread and usability of altmetrics. The usability of altmetrics indicators are concerned by many researchers, data providers and users. Only by having clear understanding of the indicators’ usability on academic evaluation, could altmetrics achieve wide acceptation and better development in the future.

In this paper, we studied the usability of altmetrics indicators in academic evaluation from different perspectives based on the data provided by Altmetric.com, in order to make objective suggestion for its usage and to perfect theoretical system.

State of research

Coverage of indicators

Coverage of an indicator means the percentage of academic work with at least count. It could reflect the probability of one piece of work to be used on the particular platform. Compared to citation, many altmetrics indicators are of short history, and are lack in unified standards or constraint as well. They don't have the ability to pay enough attention on large proportion of papers. As a result, the coverage of most altmetrics indicators is low. Among all the indicators, mendeley has the highest coverage. Hammarfelt (2014) found that mendeley has coverage of 61% on humanities-oriented papers followed by tweeters, but the coverage of many other indicators like mentioned in blogs or Facebooks is very low. Haustein, Peters and Bar-Ilan et al. (2014) found the coverage of mendeley is almost four times higher than that of citeulike, even though they are all readership of online communication platform.

For papers published in different years, the coverage they got from altmetrics indicators is different. Generally, most indicators have higher coverage on newly published papers than the old ones. Wang, Fang and Sun (2016) found the usage data to be more active on newly published papers. Cronin and Sugimoto (2014) found the mention on Twitter of hot papers spread like virus and can get to the peek in a very short time. The coverage of indicators on papers from different disciplines is also different. For papers from Art & Humanities and Life Science & Biomedicine, they could get more attention from altmetrics indicators than that from other disciplines (Costas, Zahedi & Wouters, 2015), (Holmberg & Thelwall, 2014).

Ability of reflecting papers’ impact

Many researchers doubt the value of altmetrics that it is just regarded as the popularity on social media. The related work could be summarized into three parts: correlation analysis between altmetrics and citation, predication of citation using altmetrics, and the motivation analysis of data growth of altmetrics.

A lot of studies about correlation analysis have been done based on different dataset with the similar results that mendeley is moderately correlated with citation, while mentions on social media like tweeters shows no correlation with citation, such as (Mohammadi & Thelwall, 2014), (Peters, Kraker & Lex et al., 2016), (Rosenkrantz, Ayoola & Singh et al., 2017) and (Song, 2014). Thelwall and Fairclough (2015) used mendeley to calculate JIF, and found that the JIF variants tend to be more stable, proving the value of mendeley. Altmetrics indicators could pay attention to papers more timely than citation does. If one altmetrics
indicator can be used to predict the future citation, this indicator could be regarded as a reliable metric (Thelwall, 2014). Thus some researches have been done in this way, like (Peoples, Midway & Sackett et al., 2016) on tweeters, (Shema, Bar-Ilan & Thelwall, 2014) on academic blog and (Schlögl, Gorraiz & Gumpenberger et al., 2014) on readership. Analyzing the motivation behind data increase is a method learned from text mining. Thelwall, Tsou and Weingart et al. (2013) and Na (2015) analyzed the content of tweets and found that most tweets echoed paper title or a brief summary, believing that tweets show no in-depth critical discussion of paper but they could help discover interesting researches. Jamali and Sangari (2015) found that bloggers tend to use papers to support their argument and believes that science blogs could be used to reflect papers’ value.

*Influence factors on usability*

Maflahi and Thelwall (2016) investigated the influence of time on mendeley based on papers from LIS journals, and found mendeley needs a long time to accumulate. Some researches about coverage also considered the influence of paper age or disciplines.

There are other factors like JIF, language, country, OA or not and so on. Alhoori, Ray and Kanan et al. (2015) found higher correlation with citation on papers from OA journals than that from non-OA journals. Wang, Mao and Xu (2014) found longer attention from usage data indicators on papers from OA journals. Mounce (2013) found the OA publications being able to gain more citations, and altmetrics could be used as supportive filtering system for OA in the future. There exist many advantages for developed countries in using altmetrics indicators because of the unbalanced development of social media and online communication (Moed, 2015). Many papers from developing countries are written in non-English language, or have no DOI, which are restricted from getting enough attention from international academic circles (Alperin, 2015).

*Integration of altmetrics indicatos*

Different altmetrics indicator could reflect different kind of impact. In order to establish a suitable index system and to adopt a reasonable evaluation method, multiple altmetrics indicators should be integrated. The analysis of similarities and differences between indicators was discussed in many studies.

Correlation analysis and PCA are the main methods for discovering the redundancy in value validity of indicators. Bollen, Sompel and Hagberg (2009) used PCA to analyze 39 possible factors of scholarly impact, finding that scholarly impact is multi-dimensional and cannot be measured by any single factor. Liu and Song (2016) studied data from PLOS ALM and found that correlation exist commonly among indicators, especially for usage data and online readerships. Some studies are done about indicators normalization. Bornmann believes that normalization must be done on indicators before altmetrics data to be used for measuring papers’ impact (Bornmann, 2014). Bornmann and Haunschild (2016) proposed one way to normalize tweeters by calculating normalized Twitter percentiles on journal level, and this method could be considered to be used for other indicators with low coverage. Haunschild and Bornmann (2016) used mean count on disciplines as denominator to normalize mendeley readership. Some studies are done in theoretical way. Haustein, Bowman and Costas (2015) proposed a combined altmetrics usage framework of 3 levels. Wang, Fang and Wang (2015) proposed a continuous, dynamic and comprehensive article-level evaluation system by using the indicators in different time and with different weight.
Methodology and data set

Methodology

The primary investigation contents of this paper include four main parts. Firstly we study the distribution of coverage and mean count over time for every indicator, analysing their accumulation characteristics and focusing on earlier published papers. Secondly, we study the value of indicators by correlation analysis with citation. Thirdly, correlation analysis is used to study the redundancy in value among indicators, and PCA is used to explain indicators from different dimensions. The possible affecting factors on usability are analysed then. At last, we discuss the integrated use of all indicators based on the empirical studies above.

Data set

We obtained citation data from Web of Science and altmetrics data from Altmetric.com in November 2016. We firstly choose ten disciplines from WOS and get all the journal articles published between 2006 and 2016 in those disciplines. The ten disciplines are Ecology, Biochemistry & Molecular Biology, Oncology, Neuroscience, Nursing, Geochemistry & Geophysics, Information Science & Library Science, Artificial Intelligence, Philosophy, and Educational Research. They covered five research domains of WOS. The whole number of articles is 1776187. For every article, DOI is used to collect altmetrics data from Altmetric API. Finally, we got 394678 papers with altmetrics data of 20 indicators as the main object for analyzation.

Table 1. Description for all the altmetrics indicators

<table>
<thead>
<tr>
<th>Platform type</th>
<th>Indicator name</th>
<th>Indicator meaning</th>
<th>Weight</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>online communication</td>
<td>mendeley</td>
<td>Mendeley readership</td>
<td>0</td>
<td>21.74%</td>
</tr>
<tr>
<td>communication platform</td>
<td>connotea</td>
<td>Connotea readership</td>
<td>0</td>
<td>0.43%</td>
</tr>
<tr>
<td></td>
<td>citeulike</td>
<td>CiteULike readership</td>
<td>0</td>
<td>3.05%</td>
</tr>
<tr>
<td>academic websites</td>
<td>rh</td>
<td>recommendation on F1000Prime</td>
<td>1</td>
<td>1.30%</td>
</tr>
<tr>
<td></td>
<td>wikipedia</td>
<td>citation on Wikipedia</td>
<td>3</td>
<td>1.68%</td>
</tr>
<tr>
<td></td>
<td>feeds</td>
<td>citation on academic blogs</td>
<td>5</td>
<td>2.29%</td>
</tr>
<tr>
<td></td>
<td>qna</td>
<td>mention on Stack&amp;Flow</td>
<td>0.25</td>
<td>0.05%</td>
</tr>
<tr>
<td></td>
<td>peer_review</td>
<td>peer review recommendation on Pubpeer and</td>
<td>1</td>
<td>0.28%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Publons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mainstream media</td>
<td>msm</td>
<td>mention on main social media</td>
<td>8</td>
<td>2.11%</td>
</tr>
<tr>
<td>policy documents</td>
<td>policies</td>
<td>cited by policy documents</td>
<td>3</td>
<td>0.16%</td>
</tr>
<tr>
<td>website usage</td>
<td>accounts</td>
<td>usage on Altmetric.com</td>
<td>0</td>
<td>22.22%</td>
</tr>
<tr>
<td>social media</td>
<td>tweeters</td>
<td>mention on Twitter</td>
<td>1</td>
<td>18.21%</td>
</tr>
<tr>
<td></td>
<td>fbwalls</td>
<td>mention on Facebook</td>
<td>0.25</td>
<td>4.19%</td>
</tr>
<tr>
<td></td>
<td>gplus</td>
<td>mention on Google Plus</td>
<td>1</td>
<td>0.76%</td>
</tr>
<tr>
<td></td>
<td>rds</td>
<td>mention on Rdts posts</td>
<td>0.25</td>
<td>0.29%</td>
</tr>
<tr>
<td></td>
<td>videos</td>
<td>mention on Youtube</td>
<td>0.25</td>
<td>0.11%</td>
</tr>
<tr>
<td></td>
<td>weibo</td>
<td>mention on Sina Weibo</td>
<td>1</td>
<td>0.07%</td>
</tr>
<tr>
<td></td>
<td>pinners</td>
<td>mention on Pinterest</td>
<td>0.25</td>
<td>0.01%</td>
</tr>
<tr>
<td></td>
<td>linkedin</td>
<td>mention on LinkedIn</td>
<td>0.5</td>
<td>0.01%</td>
</tr>
<tr>
<td></td>
<td>book_reviews</td>
<td>mention on Syllabus</td>
<td>1</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

Altmetric.com provides an integrated score (Altmetric Attention Score, in the following text, we will use “Score” instead) for every paper by calculating all indicators with different
weight. As seen in table 1, mendeley, tweeters and accounts have higher coverage than other indicators. The coverage of Pinners, linkedin and book_reviews is very low, so we don’t analyse them in the following part.

The rows with dark shading are the indicators with zero weight for Altmetric Attention Score. When grasping data from Altmetric API, if one paper’s Score is zero, nothing would return but “null”. In other studies like (Haustein, Costas & Larivière, 2015), the coverage of mendeley is around 60-80%, but based on our dataset it is only 21.74%. This happens because when grasping data from Altmetric API, many papers got zero in Score while their mendeley readership is not zero, so “null” returned and the mendeley readership of that paper is not observed. The same happens to connotea and citeulike.

**Procedures**

**Distribution of coverage and mean count over time**

We analyzed the coverage distribution of Score over time and disciplines in Figure 1 to show the overall coverage trend for indicators. We find that for all papers of any discipline, the coverage increases over time, especially after the year of 2011. The coverage on papers published earlier than 2011 is quite low, but larger than zero. This phenomenon indicates that altmetrics indicators not only pay attention on newly published papers, but also on old ones.

Papers from the research domain of Life Sciences & Biomedicine have higher coverage. This kind of distribution over disciplines is suitable for most indicators except for some with obvious aptness on particular discipline. The platform of rh only evaluates and recommends papers of Life Sciences & Biomedicine. As the platform of quora, Stack & Flow is the website discussing IT problems.

![Figure 1. The coverage distribution of Altmetric Attention Score over time and disciplines](image)

In order to analyze their trend and connection, We calculate the coverage and mean count of every indicator over time in Figure 2. For most indicators, their coverage starts to increase after 2010. Some platforms start to grow or prosper since 2010, so the papers published after 2010 would receive more attention. The coverage of rh, wikipedia, policies keep decreasing from 2006. Their platforms have already been mature, so they would focus on the literature that has been recognized by the academia and the society. Different from coverage, the mean count of indicators show three main trends, including increasing, decreasing and keeping stable, in Figure 2.

Comparing the trend of coverage and mean count for every indicator, we could analyse their attention to newly or earlier published papers. Some indicators perform continuously while some others just pay attention to newly published papers. For example, mendeley has the increasing coverage, while mean count decreasing over time, which means that the newly published papers could get attention there but still need time to accumulate. Connotea and citeulike from online communication platform, feeds and peer_review from academic websites show the same trend. But wikipedia, rh and policies show a decrease in both coverage and mean count, which means that old papers that have been tested are more easily to get attention on those platforms. For indicators from social media, mainstream media and
usage data, their coverage and mean count keep increasing, for the reason that newly published papers could attract attention but this kind of attention doesn’t last long.

Figure 2. Indicators’ coverage and mean count on papers of different published year

Value of every indicator

Correlation analysis is the most commonly used method in proving indicators’ value. Researchers believe that correlation existing between altmetrics and citation could prove the indicators’ ability to reflect papers’ impact. Our study use Spearman Correlation to analyse the relation between every altmetrics indicator and citation. The result is shown in Figure 3. We found that indicators from the same type of platforms tend to have similar correlation coefficient. Most indicators from online communication platform and academic websites show low to medium correlation with citation. Mendeley has the highest coefficient of 0.66. But other indicators like social media mention or usage type have no correlation with citation.

Figure 3. Correlation between altmetrics indicators and citation

In order to make comprehensive understanding of the correlation between altmetrics and citation, we analysed the distinctions of correlation based on papers of different disciplines and different ages.
Figure 4 shows the changes of the correlation coefficient over time. As citations need at least 3 years to accumulate, we used papers published before 2013 for analysis. The correlation of mendeley with citation reduced after the year of 2010, which means that for papers newly published, mendeley has the space to increase and mendeley could reflect the value of earlier published papers to some extent. The same phenomenon was observed from citeulike, connotea and wikipedia. For rh and feeds, their correlation with citation keep increasing till 2012, after which it begins to fall, indicating that these indicators need time to increase but their ability of discovering old papers is not as strong as citation. Other indicators like tweeters and accounts keep increasing till 2013, proving their weak ability of reflecting the value for newly published papers.

For papers from different disciplines, the correlation between altmetrics indicators and citation shows no obvious distinction, except for citeulike which has lower correlation in Natural Science, and rh which has much higher correlation in Life Sciences & Biomedicine.

**Redundancy of indicators’ value**

We found low to medium correlation existing commonly between altmetrics indicators. Indicators from same type platforms have similar meaning of data, which makes them easily to have correlation with each other.
Accounts has correlation with many indicators, especially those from social media platform. It has high correlation with tweeters, and low correlation with feeds, msm, fbwalls and gplus. Mendeley is another indicator which has correlation with many indicators. Those indicators are citeulike, rh, feeds, msm, accounts and gplus, most of which are from online communication platform and academic websites. Besides, low correlation exists between many other indicators as shown in Figure 5.

For indicators from platform of same type: mendeley, connotea and citeulike, all of which are from online communication platform, have low correlation with each other; tweeters and fbwalls have low correlation, while other indicators from social media show no correlation. For indicators from different type of platform: most indicators from online communication platform have correlation with that from academic websites; indicators of usage have correlation with most indicators from social media platform and mainstream media; msm has correlation with many indicators from social media platform.

In order to find the reason behind redundancy, we use PCA (Principal Component Analysis) to analyze the indicators from different dimensions. Citation is included for comparison. The first two components could explain 70% of the variance, while the following components keep score around 1, indicating that they also represent information of different dimensions and cannot be ignored. All the indicators variances on principal component are shown in coordinate in Figure 6.

Based on the variance of every indicator on components, we could summarize the two components as follows: PC1 means timeliness of indicator, and the negative direction means indicator could reach the top just after a short time; PC2 means the aspect of value indicators could reflect, and the positive direction means the indicator tends to reflect the academic value of papers, while the negative direction means the indicator tends to reflect the social value.

According to Figure 6, we could see three obvious groups of indicators. The first group includes z9, mendeley and citeulike, which tend to reflect the academic value of papers and need a period of time to accumulate. Besides the three, other indicators like rh, connote show the similar character with lower degree. The second group includes accounts and tweeters, tending to reflect the social value of papers timely, and fbwalls, msm, rdts and gplus have the same character but lower degree. The third group is around the origin point, with the meaning of no obvious tendency to both PCs. Their character may be explained in other dimensions. PCA could just give rough explain on character tendency.

![Figure 6. The distribution of altmetrics indicators on Principal Components](image-url)
Analysis of influence factors on usability of altmetrics

We assume that the journal impact factor is one of the factors. The total number of journals having at least one paper with Score is 1900, and the top 10 (sorted by the whole number of papers with Score) of them are all from Life Sciences & Biomedicine. We use JIF, 5JIF and NE as quantitative indicators to represent journal impact, which are published in July 2016. 5JIF refers to the mean JIF of recent five years. NE, short of Normalized Eigen factor, is released as a supplement for JCR for cross-field comparison. The NE of one journal is 2, which means this journal impact is twice than the average impact of journals from the same discipline. We found more than 71.57% of papers is with NE larger than 1, and the correlation coefficient of NE and Score is 0.173 on these papers. On the remaining part, the coefficient is 0. We could get the point that altmetrics pay more attention to papers from journals of high impact.

In order to study the influence on every single altmetrics indicator, we did correlation analysis between altmetrics indicators and the three journal impact indicators, including citation. We got the result that for indicators like accounts and msm, the correlation with JIF and 5JIF is higher than that with citation, indicating that the indicators with short accumulation period would be more easily affected by journal impact. Comparatively, indicators like mendeley, Wikipedia, citeulike and policies which have long accumulation period would be more easily affected by citation than by journal impact.

Figure 7. Correlation between altmetrics indicators with journal impact indicators

Altmetrics indicators have prejudice against non-English papers. English is the main language used in online communication platform, as well as the social media and academic websites. In order to make a easy communication with others in the international academic circle, researchers tend to use English to write papers. For all papers got from WOS, 34 types of language are used, and the percentage of English papers is 97.67%. For papers with Score, 10 types of language are used, and the percentage of English papers is 99.73%.

The prejudice also exists in papers from developing countries, mainly because of the insufficient use of DOI and the non-English language used. We firstly analyze the DOI distribution all over the world to have a glimpse of DOI popularity for different areas. We found that for countries from North America, Europe, Oceania and East Asia, DOI is more widely used than that of Africa, South America and Middle Asia. Then we use the first author’s nation to represent the papers’ source country, to analyze the coverage of altmetrics all over the world. The result is shown in Figure 8. The top 10 countries with highest percentage are Switzerland, USA, UK, Denmark, Ireland, Austria, Netherlands, Sweden, Italy and Germany, all of which are from developed areas. For some countries with low percentage, most papers are written in non-English languages and have quite low DOI popularity. For example, 63.81% of all papers from Russia are written in Russian and only 10.44% papers have DOI.
We also assume that collaboration on author level and country level would have an influence on the usability of altmetrics. 73.79% of papers with Score are written by multiple authors, while the percentage of all papers from WOS is 56.02%. With the data of country that we collect from authors’ address, we found that 26.92% of papers with Score are internationally collaborated, while the percentage for all papers from WOS is 18.45%. This shows that multiple authors collaborated papers and internationally collaborated papers could both get more attention from altmetrics. Analyzing on papers of different published year shows that the percentage gap of international collaboration between all papers and papers with Score is reducing over time. Few differences have shown for different indicators.

![Figure 8. The percentage of papers with Altmetric Attention Score all over the world](image)

**Discussion**

The coverage and mean count of most altmetrics indicators change over time. Indicators from online communication platform and academic websites need time to accumulate. Earlier published papers could also get attention from these indicators. Indicators from social media, mainstream media tend to pay attention to newly published papers. Some indicators from platforms that have got enough development tend to pay attention to papers that have been tested by time. For different indicators, the moment they pay attention to and the time to accumulate are different.

For the correlation with citation, mendeley has medium coefficient, other indicators from online communication platform and academic websites have low coefficient, while the coefficient for indicators from social media, mainstream media and usage data is around zero. By correlation analysis among all indicators, we found redundancy exist commonly in value for indicators from same type platforms and different type platforms. Mendeley and accounts are the two indicators that have most commonly correlation with other indicators. By doing PCA, we conclude two main components, timeliness and value impact, to explain indicators. But as other components having high percentage on variance, we don’t regard the PCA having the good result as usual, believing that some information would miss if doing so. This conclusion supports the diversity of indicators.

We found factors like disciplines, JIF, languages and DOI popularity could influence indicators’ usability. For example, papers from Life Sciences and Biomedical Sciences could get more attention from many indicators, in-time indicators such as tweeters would suffer from the impact of source journal, and non-English papers or papers from undeveloped areas would get less attention they deserved. For the effect of OA, we did not get the chance to analysis because of the dataset we got (all the papers are from non-OA journals).

The differences in accumulation rate, the attention for earlier published papers, the redundancy and diversity in value, make it hard to integrate use of indicators. All these
problems must be considered, as well as the better performance of indicators on English papers, papers from developed areas, and papers from Life Sciences & Biomedicine. The normalization for indicators which is one of essential parts (Bornmann, 2014) is just at the exploring period, with mendeley and tweeters being the researching objects and normalization over time and disciplines being the main method. Many things need to be done before a universal framework proposed. The integrated use of indicators needs guides in theory as well as supports from practice.

Conclusions

In this article, we studied the usability of altmetrics indicators on academic evaluation from many perspectives. The study was carried out from the count level to the value level of indicators, with the method of comparative analysis, correlation analysis and PCA. By analyzing the distribution of coverage and mean count over time, we studied the accumulation character for every indicator and their attention to earlier published papers. By analyzing correlation between citation and indicator, and among indicators, we studied the value and the redundancy on value for every indicator. By doing PCA for indicators, we explained the indicators on different dimensions, and have better understanding on the diversity and aptness of indicators. We also analyzed possible factors that may make influence on the usability. At last, we discussed the integrated use of indicators, and found the research of usability and universal framework still has a long way to go.

Acknowledgments

We gratefully acknowledge support from ISTIC-EBSCO Joint Laboratory for Big Data Discovery Service.

References


Identifying and tracking scientific and technological knowledge memes from citation networks of publications and patents

Xiaoling Sun\textsuperscript{1} and Kun Ding\textsuperscript{2}

\textsuperscript{1}xlsun@dlut.edu.cn
Dalian University of Technology, Dalian (China)

\textsuperscript{2}dingk@dlut.edu.cn
Dalian University of Technology, Dalian (China)

Abstract
Science and technology are becoming more and more closely linked today and the research on the relationship between them has become a hot topic. In this paper, we are trying to investigate the relationship between science and technology at micro-level. A new carbon nanomaterial - graphene is taken as an example, and publications and patents are used as data sources for the representation of science and technology. Citation networks of publications and patents are constructed, on which a knowledge meme discovery algorithm is used, in order to identify scientific knowledge memes and technological knowledge memes that play a key role in the evolution of scientific and technological knowledge. We also track the diffusion and co-occurrence of knowledge memes, which could provide guidance for promoting scientific and technological innovation and making research policy.

Conference Topic
Citation and co-citation analysis; Patent analysis

Introduction
The relationships between science and technology are complex, and a number of endeavours have been undertaken to reveal the connection and interaction between science and technology in recent years (Meyer, 2001; Van Looy et al., 2003; Boyack and Klavans, 2008; Narin and Noma, 2005; Zhao and Guan, 2012, 2013; Huang et al., 2015). Some researchers find science and technology are becoming more and more closely linked today (Meyer et al., 2010). However, the relationships are often different in different fields and at different phases of a technological “life cycle” (Brooks, 1994).

Publications are the main channel for documentation of scientific findings, and patents reflect the development of new technologies. Bibliometric methods based on paper and patent analysis are the main methods to study the relationships between science and technology. These methods are often based on the citation of papers and patents (Chen and Hicks, 2004; Narin and Noma, 2005; Yeh et al., 2013; Huang et al., 2015), or the relationship between the inventors of the patent and the authors of the paper (Meyer, 2006; Wang and Guan, 2010). This “science-technology relationship” research based on these methods mostly reveal the quantitative characteristics of the relationship, however, the research on the interrelated topics between science and technology is still relatively rare.

Graphene - a new type of carbon nanomaterials is taken as an exemplary research field in this study. It has attracted the attention of a large number of researchers and has been widely used to modify various polymers in recent years (Novoselov et al., 2005; Geim and Novoselov, 2007). The potential applications in lithium-ion batteries, solar cells, sensors and supercapacitors have been extensively studied and have great application prospects (Wu et al., 2010). Over the past few years, graphene/polymer nanocomposites have made encouraging progress in basic research and product application development, but from basic research to extensive application there is still a long and complex process. With the “Material Genome Project” proposed, how
to accelerate the process of research and application of materials to promote the integration of science and technology has become the common needs of countries around the world. Therefore, the development of effective methods to explore the relationship between science and technology, and promote their integration is a key issue. Although this paper focuses on the field of graphene, the related methods can be extended to other fields.

In this paper, we are going to investigate the relationship between science and technology at micro-level, in which “knowledge memes” are identified and tracked. Dawkins introduced the term “meme” (Dawkins, 1976) for the entities such as words, ideas in cultural evolution that are similar as genes. Memes are the cultural equivalent of genes that spread across human culture by means of imitation (Kuhn et al., 2014). Here we use also use “meme” to refer to the knowledge that plays an important role in shaping the evolution of science and technology. Based on the knowledge memes, this study offers a further description of knowledge diffusion path and knowledge innovation, with the help of meme co-occurrence networks.

The rest of this paper is organized as follows: Section 2 reviews related work; Section 3 introduces the proposed method including how to identify knowledge memes and build knowledge meme networks; Section 4 presents the results of the proposed method; then we conclude in Section 5.

Related Work

The study of the relationship between science and technology can be traced back to the study of Price in the 1960s (Price, 1965). He believed that science and technology had their unique knowledge accumulation structure. When the paradigm is broken, knowledge may flow from science to technology or flow from technology to science. Bibliometric methods based on publication and patent analysis are the main methods of studying the relationship between science and technology. Commonly used data comes from USPTO, EPO and other patent offices and Derwent Innovation Index (DII) database for patent data and Web of Science (WoS) for publication data. The main methods can be divided into three categories: patent’s paper citation analysis, paper’s patent citation analysis, “inventor - author linkage” analysis.

One approach of exploring science-technology relationships consists of taking into account citations to scientific papers found in patents, which reveals the promotion of basic science research to technological innovation. Patent citations, particularly the citations to non-patent references (NPRs) (Verbeek et al., 2002), are regarded as the most popular indicator to track the linkage between science and technology (Chen and Hicks, 2004). Narin and Noma (Narin and Noma, 2005) analysed the citation relationship between patents and papers in the field of bioscience, and concluded that patents and papers were highly correlated, and high-tech was closely related to science. Meyer (Meyer, 2001) used patent’s paper citation information to analyse the relationship between nanoscience and nanotechnology. The linkage established through the reference field of patents or publications, however, are doubtful, mainly because of different citation behaviours and different citation motivations among authors, inventors and patent examiners (Meyer, 2006). The above studies use citations to papers in patents for analysis, observing only one aspect of the interaction.

The citation of patents in academic papers also contains the relationship between basic research and technological innovation. Glanzel and Meyer (Glanzel and Meyer, 2003) analysed the similarities and differences of different disciplines based on the citation of patents in SCI papers, and found that the chemical field had cited more patents than other disciplines, and the patents
on chemistry, medicine and medicine were cited more frequently. Some scholars combined the patent’s paper citation analysis with paper’s patent citation analysis to reveal the two-way interaction between science and technology (Huang et al., 2015).

Furthermore, some studies quantified the interaction between science and technology through inventor-author linkages (Meyer, 2006; Wang and Guan, 2010) by tracking the patent inventors published research papers or researchers invented patents. Alternative approaches concerned about how to explore joint publications between academia and industry (Van Looy et al., 2003, 2006). Similarly, Zhao and Guan (Zhao and Guan, 2012, 2013) qualified and evaluated the associations between scientific activity and technological output in the case of innovation system.

Most of the researches have analysed the quantitative characteristics of the relationship between science and technology, but the research on the interrelated topics between science and technology is still rare. In the development of science and technology, a key knowledge element may play the role of “knowledge meme”, which determines the evolution and mutation of specific domain knowledge. The identification and research of knowledge memes may shed light on the understanding of the development of the research field. Recent research on memes is interested in meme diffusion in social media, such as blog space and Twitter (Ratkiewicz et al., 2011). The issue of competition among memes with limited attention also draws researchers’ attention (Weng et al., 2012). In science, the spread of a knowledge meme can be tracked via citations. Kuhn et al. (Kuhn et al., 2014) proposed a method based on citation networks to extract scientific memes in the scientific literature. They also give a definition: A scientific meme is a short unit of text in a publication that is replicated in citing publications.

In this paper, we follow the definition of Kuhn et al., and also extend it to technological meme: A technological meme is a short unit of text in a patent that is replicated in citing patents. The identification of knowledge memes could help us to understand the evolution and interaction of science and technology more deeply from the micro level. Furthermore, the exploration and the recombination of knowledge memes, could promote the development and innovation of science and technology.

Materials and Methods

Dataset
The study exploits a patent database of graphene research field from Derwent Innovations Index (DII). The patents that have “graphene” in title, keyword or abstract are included in the patent database. Publication data related to graphene is derived from Web of Science (WoS) based on the same search strategy and stored in the publication database. The data covers the period 2010-2015 and there are totally 15,818 patents and 71,008 publications.

Building paper and patent citation networks
The inheritance of knowledge by citation processes reflects the evolution of knowledge. When paper cites the previous papers, the part of knowledge that is similar has been retained. The most relevant and most valuable knowledge has been inherited, which becomes the knowledge memes. Thus, the citation relationship provides an advantageous resource for the discovery of knowledge memes.

The main storage medium of scientific knowledge is papers. The citation relationships between papers show the communications of vertical inheritance and horizontal association of scientific
literatures. They reflect the trajectory of scientific development and the inheritance of knowledge. Similarly, the main storage medium of technological knowledge is patents, and there are also citation relationships between patents.

In this paper, we use the citation relationships to build paper citation network and patent citation network. The nodes in the network are papers (or patents), and the links between them represent the citation relationships.

**Identifying knowledge memes**

The identification of scientific knowledge memes and technological knowledge memes can help to discover what knowledge memes play a key role in the evolution and mutation of domain knowledge and predict the future development trend of the field.

Words or phrases that often occur in the literature can be considered as important memes, but many frequent words such as “method” are of little significance to a field. Kuhn et al. proposed a scientific memes identification algorithm based on the citation network, mainly using the frequency of occurrence and the degree to which they propagated along the citation graph. A meme that is often appear in publications that cite meme-carrying publications but rarely appear in publications that do not cite a publication that already contains the meme (Kuhn et al., 2014), has a higher score. $P_m$ is defined as the propagation score:

$$P_m = \frac{d_{\rightarrow m}}{d_{\rightarrow m}}$$  \hspace{1cm} (1)

where $d_{\rightarrow m}$ is the number of publications that contain meme $m$ and cite at least a publication that contains $m$; $d_{\rightarrow m}$ is the number of publications that cite at least a publication that contains $m$; $d_{\rightarrow m}$ is the number of publications that contain meme $m$ but did not cite any publications that contain $m$; $d_{\rightarrow m}$ is the number of publications that did not cite any publications contain $m$.

The method also considers the frequency of the meme $m$, using $f_m$ to represent the proportion of the publications that contain $m$. Finally, the score of meme $m$ is defined as:

$$M_m = f_m P_m$$  \hspace{1cm} (2)

The propagation score, as defined in Eq. (1), could be improved by adding a small amount of controlled noise $\delta$, which we will try in the future work.

The words in titles and keywords after removing stop words are used as candidates for memes. Based on this method, this paper extents it to the paper and patent citation networks, discovering scientific knowledge memes and technological knowledge memes at the same time, and further investigating the relationships between science and technology.

**Constructing knowledge meme co-occurrence network**

On the basis of the identification of knowledge memes, we construct a knowledge network, which is a network with knowledge memes as nodes and the co-occurrence relationships among memes as edges. The weight of the edge is the number of times two memes appear together. Based on the knowledge meme network, the community discovery algorithm (Blondel et al., 2008) is used to identify closely related meme subgroups. The knowledge memes in a community have a close relationship, the recombination of which may reveal possible hidden knowledge innovation.
Results

Scientific memes and technological memes

All the words in the title and keywords after removing the stop words of the papers and patents are considered as potential memes. After building the paper citation network and patent citation network, we use Eq. 2 to compute the scores of the memes. A higher score indicates the word or phrase is more likely to be a meme in the research field. Table 1 and 2 show the top 10 scientific memes and technological memes in the year 2010, 2012 and 2014. The annual data used for the calculation of memes is the cumulative data from 2010 to the current year.

It could be seen from table 1 that the scientific knowledge memes such as Metamaterial, Nanocomposite, OLED and Laser have been the common memes in graphene research field and have been playing an important role in the development of the field. As time goes by and the progress of technology, knowledge memes are also evolving. The identification of these memes is of great importance in understanding the evolution of the field.

From table 2, we could see that the technological memes are more about the preparation of materials and the research of process performance. Polymer, Composite, Substrate, Film, Electrode, OLED, etc. have always been the knowledge memes in graphene technological field. In 2014, the focus of technological research is still similar with that of a few years ago, there is no obvious evolution in the topics.

<p>| Table 1. Top 10 scientific memes and their scores in the year 2010, 2012 and 2014. |
|---------------------------------------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|</p>
<table>
<thead>
<tr>
<th></th>
<th>2010</th>
<th>2012</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphene</td>
<td>1.141</td>
<td>0.984</td>
<td>0.925</td>
</tr>
<tr>
<td>Metamaterial</td>
<td>0.554</td>
<td>0.538</td>
<td>0.472</td>
</tr>
<tr>
<td>Nanocomposite</td>
<td>0.513</td>
<td>0.389</td>
<td>0.398</td>
</tr>
<tr>
<td>Nitrogen</td>
<td>0.466</td>
<td>0.386</td>
<td>0.397</td>
</tr>
<tr>
<td>Nanoco</td>
<td>0.457</td>
<td>0.353</td>
<td>0.336</td>
</tr>
<tr>
<td>Nanoparticle</td>
<td>0.439</td>
<td>0.351</td>
<td>0.334</td>
</tr>
<tr>
<td>OLED</td>
<td>0.438</td>
<td>0.314</td>
<td>0.328</td>
</tr>
<tr>
<td>Polyaniline</td>
<td>0.399</td>
<td>0.305</td>
<td>0.316</td>
</tr>
<tr>
<td>Nanoribbon</td>
<td>0.363</td>
<td>0.289</td>
<td>0.296</td>
</tr>
<tr>
<td>Quantum dot</td>
<td>0.332</td>
<td>0.288</td>
<td>0.296</td>
</tr>
</tbody>
</table>

<p>| Table 2. Top 10 technological memes and their scores in the year 2010, 2012 and 2014. |
|---------------------------------------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|</p>
<table>
<thead>
<tr>
<th></th>
<th>2010</th>
<th>2012</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polymer</td>
<td>0.813</td>
<td>0.823</td>
<td>0.892</td>
</tr>
<tr>
<td>Graphene</td>
<td>0.774</td>
<td>0.784</td>
<td>0.850</td>
</tr>
<tr>
<td>OLED</td>
<td>0.729</td>
<td>0.617</td>
<td>0.797</td>
</tr>
<tr>
<td>Carbon</td>
<td>0.641</td>
<td>0.521</td>
<td>0.647</td>
</tr>
<tr>
<td>Material</td>
<td>0.582</td>
<td>0.503</td>
<td>0.604</td>
</tr>
<tr>
<td>Comprise</td>
<td>0.556</td>
<td>0.494</td>
<td>0.599</td>
</tr>
<tr>
<td>Layer</td>
<td>0.513</td>
<td>0.494</td>
<td>0.574</td>
</tr>
<tr>
<td>Involve</td>
<td>0.449</td>
<td>0.482</td>
<td>0.573</td>
</tr>
<tr>
<td>Composite</td>
<td>0.444</td>
<td>0.476</td>
<td>0.570</td>
</tr>
<tr>
<td>Prepare</td>
<td>0.437</td>
<td>0.475</td>
<td>0.556</td>
</tr>
</tbody>
</table>
Relationships between graphene science and technology based on knowledge memes

With the help of the identification of scientific knowledge memes and technological knowledge memes, we could investigate the relationships between graphene science and technology from the view of knowledge memes as follows:

**Differences:** Graphene scientific research is mainly carried out from three aspects: the preparation of graphene, graphene applications, and graphene composites. The range of applications is wide, in particular, including batteries, superconductors, semiconductors, capacitors, carbon nanotubes, Lasers, cell imaging, etc. Graphene technology is mainly focused on the preparation of graphene and part of graphene applications, such as the preparation of graphene composites, films, electrodes, lasers, which is more concentrated.

**Similarities:** The preparation of graphene is the focus of both science and technology, and in the applications, the battery, laser, composite materials are also important common concerns. In these aspects, the interaction between science and technology is frequent and it could be easy to apply the findings of basic research to application. In other aspects, there needs more efforts to promote the two-way interactions.

**Development Trends:** Although there are some differences in the research topics between graphene science and technology, the relationship between them is becoming more and more close. The boundary is becoming more and more blur, showing convergence trend.

Some researchers found that the time between the publication of a journal article and the patent application citing the article was relatively short (Carpenter et al., 1980). In Fig.1, we plot the time spans between the publication year of patents and the average publication year of the cited papers in the field of graphene. We obtain similar results: the time span is showing a shrinking trend. This may indicate that the knowledge flow and the transformation from basic research to application is becoming quicker.

![Figure 1. Publication year of the patents vs. publication year of the cited papers (the numbers in the figure are the average publication years of the papers, and the numbers on the right are the time spans between the publication year of the patents and the average publication year of the cited papers).](image)

We also examine the similarity between the titles of patents and papers in order to identify a consensus in the use of terms between science and technology. The titles of patent and paper are represented as two vectors: \( \mathbf{A} \) and \( \mathbf{B} \). \( \mathbf{A} \) and \( \mathbf{B} \) have the same dimension (the combination of terms in the titles of patents and papers), and the components of the vectors are 1 if the titles
of the patent or the paper have the corresponding words. Then, the similarity between two vectors $\mathbf{A}$ and $\mathbf{B}$ is computed using cosine similarity:

$$
similarity(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \tag{3}
$$

where $A_i$ and $B_i$ are components of vector $\mathbf{A}$ and $\mathbf{B}$ respectively.

In a certain year, the similarity between the title of a patent and the aggregated titles of its cited papers is computed using cosine similarity, and then the topic similarity is averaged using the number of patents in that year. Fig. 2 plots the topic similarity between patents and the cited papers in subsequent years. In general, the average similarity is increasing from the year 2010 to 2015, although not very obvious.

![Figure 2. Topic similarity between patents and the cited papers (the numbers in the figure are the average similarities per year).](image)

**Knowledge meme diffusion path on citation networks**

Based on the identification of knowledge memes, we could track the meme diffusion path on citation networks. For example, in Fig. 3(a), which is the patent citation network with memes, we could see: a. the key patents that play an important role in the knowledge meme diffusion; b. the diffusion path of knowledge memes. Some patents play as the basis of other patents, such as patent “US7071258-B1”, the title of which is “Nano-scaled graphene plate material used for composite material, comprises nanometer-scaled plate(s) comprising sheet(s) of graphite plane, and length, width and thickness of nanometer scaled plates are specified”. The memes it contains are “composite”, “graphite”, “comprise” and “material”. It provides the basic knowledge about nano-scale graphene material which is useful for other patents. Meanwhile, some patents, for example, patent “US7790242-B1” (Electrostatic deposition of graphene on substrate involves applying voltage across graphite sample secured to positive electrode and substrate placed over a second electrode, so as to transfer graphene from graphite sample to the substrate), inherits the knowledge from many other patents, the memes of which are “involve”, “oled” and “graphite”. Different patents belong to different levels in the development of technology.

Similarly, scientific knowledge meme diffusion path could also be tracked as seen from Fig. 3(b). For example, we mark two papers that on the diffusion path of the meme “OLED”, paper (DOI: 10.1021/nn1015874: “Organic Photovoltaic Devices Using Highly Flexible Reduced Graphene...
Oxide Films as Transparent Electrodes”) and paper (DOI: 10.1002/pola.23802: “Thermosensitive Graphene Nanocomposites Formed Using Pyrene-Terminal Polymers Made by RAFT Polymerization”). The paper (DOI: 10.1002/pola.23802) contains memes “oled”, “nanoco” and “nanocomposite”, and inherits the knowledge from many other papers.

Figure 3. Knowledge meme diffusion (a) on the patent citation network in 2010 (nodes are patents with memes) (b) on the paper citation network in 2010 (nodes are papers with memes). Different colors indicate different clusters of patents identified by community detection algorithm.

Knowledge innovation based on meme co-occurrence networks

In this section, knowledge meme networks are built based on the co-occurrence of memes, from which we could get insight into the knowledge memes that are always appear together, and explore possible knowledge innovation based on the recombination of knowledge memes.

Fig.4 shows an example of technological knowledge meme co-occurrence networks in the year 2010 and 2012. “Graphene” is ignored as it almost occurs with all the other memes, and affects the results. In 2010, the memes are basically formed a large group and closely related, which is mainly about the preparation of materials. In 2012, using community detection algorithm, there are three groups. The memes contained in each group are: dioxide, oxide, titanium, involve, fiber, saturate, oled, laser, graphite, prepare; layer, film, memory, device, semiconductor, form, transistor, substrate, carbide, solar, electrode, silicon; composite, carbon nanotube, battery, lithium, carbon, material, nanotube, resin, phosphate, comprise.

Figure 4. Technological knowledge meme co-occurrence networks in 2010 (a) and 2012 (b).
Fig. 5 is the scientific knowledge meme co-occurrence networks in the year 2010 and 2012. In 2010, there are several memes closely related, e.g. oled, nanoparticle, nanocomposite and nanoco. In 2012, we could see the structure is very different from that of patent network. The three groups are: nanocomposite, nanoco; graphene, quantum well, molecular beam, nanoribbon, isotope-separation, hydrogen-peroxide detection, nanopore, nanoparticle, diffusion dynamics, superconductor, silicon, nanofluid, single dirac cone, metamaterial, organic conductor, graphene quantum dot, graphyne, oled, single-layer mos2; q-switch, mode-lock, fiber laser, saturable absorber, laser.

![Figure 5. Scientific knowledge meme co-occurrence networks in 2010 (a) and 2012 (b).](image)

The memes that are in one group that have not linked together, may have the chance of recombination to give rise to knowledge innovation. In the future work, we would like to apply link prediction algorithm to predict potential links between knowledge memes.

Conclusions
In this paper, the knowledge meme identification algorithm is used to identify the scientific and technological knowledge memes, and the knowledge meme network is constructed, which is applied to the graphene field to investigate the relationship between science and technology and track the diffusion of memes. The results reveal the similarities, differences and development trend between science and technology in the graphene field, and the related research can be extended to other fields and provide guidance for making science and technology policy and promoting technology innovation.

This study is a first attempt to investigate the relationship between science and technology in the view of knowledge memes at micro-level. There are also some limitations of this study. Firstly, the relationship between science and technology is not fully explored, in order to solve this, we are building citation networks between papers and patents to fully link science and technology. Secondly, based on the knowledge meme networks, we will try to use link prediction algorithms to study the recombination of memes, model the association trend between scientific and technological memes, making efforts to accelerate the transformation of science to technology and the promotion of technology to science.

Acknowledgments
This work is partially supported by grant from the Postdoctoral Science Foundation of China (No.2015M581337, No.2016T90224), the Fundamental Research Funds for the Central Universities, Open Fund Project of ISTIC-THOMSON REUTERS Joint Lab for Scientometrics Research.
References


Event Detection in Scientific Mapping based on a Novel Structural Community Similarity Algorithm

Meng Xiangfeng1 Liu Xinhai1 Zhang Yan1 Glanzel Wolfgang2

1 pku.ericmeng@foxmail.com, xinhai.liu@yahoo.com, zhy@cis.pku.edu.cn
Peking University, Beijing, China

2 wolfgang.glanzel@kuleuven.be
Hungarian Academy of Sciences, IRPS, Hungary

Abstract
Science is a dynamic system. Its cognitive structure is subject to permanent changes. Any particular structural change, such as new research trends and emerging topics, can be detected in cognitive maps. In this paper, we attempt to identify new research trends and topic transition by detecting both changing structure and anomalous events in the course of mapping the structure of journal-based clustering. We form one community for each journal cluster, with nodes representing journals and edges representing the citation links among journals. To measure the significance of changes in the dynamic journal clusters, a community similarity measurement is required. However, the existing similarity algorithms neglect the shift of the internal structure in networks and graphs, which determines that these algorithms cannot be directly applied to journal clusters. We propose a novel community-similarity algorithm, which considers both the shifts of vertices and the shifts of communities’ layered structure. Communities’ layered structure categorizes nodes into different groups, depending on their influence on the community. We apply the novel algorithm on the temporal journal data set, and identify two types of anomalous events. Both the visualizations of the journal clusters and the text annotations demonstrate that the identified events correspond to new emerging trends in scientific fields.

Conference Topic
Knowledge discovery and data mining, Scientometrics

Introduction
Mapping of science has played an important role in the field of library and information science. It helps improve the subject delineation and identify the emerging and converging fields (Janssens, Zhang & De Moor, 2009). However, science is a dynamic system. New research topics emerge, others might disappear and increasing inter- and cross-disciplinarity also affects structure and relationships of the concerned disciplines. Not all of those can be predicted or follow linear evolitional paths. Breakthroughs, delayed recognition, cross-fertilization may lead to unexpected changed in the network of scholarly communication. Finally, a critical mass of research results and literature is needed to make changes visible (Glanzel & Thijs, 2012). These can be regarded as anomalous events that occur in the development of science mapping.

Journal clustering can uncover such changes and provide insights into the emergence of new research topics (Liu, Glanzel & De Moor, 2012; Glenisson, Glanzel & Janssens, 2005). Previous studies have proven that journal clustering can be employed as a good reference for journal categorization (Liu, Yu & Janssens, 2010). It is also efficient to be applied to detect new trends in different scientific fields. Thus in this paper, we adopt the temporal journal data set from Clarivate Analytics Web of Science Core Collection, and attempt to detect anomalous events that occurred in scientific mapping.

We analyze the temporal journal data set based on their citation relations. Dynamic journal networks are constructed, with node representing journals and edges representing the direct citations among journals. Journal clusters can be obtained by partitioning the journal networks. Particular, journal clusters can be regarded as communities. Thus, to detect the anomalous
events in journal clusters, a notion of community similarity is required to measure the significance of changes in the evolution of these temporal journal clusters. Several studies have focused on algorithms that measure the similarity of two graphs or networks. Without loss of generality, we do not distinguish networks from graphs in this paper. Similarly to community detection, graph similarity algorithms are able to track temporal networks (Palla, Barabasi & Vicsek, 2007), to monitor the evolution of networks (Koutra, Vogelstein & Faloutsos, 2013), notably to track the changes over time and to detect anomalies (Noble & Cook, 2003) and events (Caceres, Berger-Wolf & Grossman, 2011). Multiple graph similarity algorithms have been proposed by both Bunke (2007) and Papadimitriou (2010). However, the existing algorithms ignore the structural shift of communities. Thus, the existing algorithms are not suited to identify communities with largely changing structure. For instance, the shift of the core journals in one journal cluster indicates that potential topic transition may have occurred. Therefore, approaches that are more convincing are needed to identify the shift of such structure.

This paper presents a novel community similarity algorithm by considering both the shift of communities' vertex set and the changes in communities' structure. Communities' structure determines the way in which individuals in communities are organized. It categorizes nodes into different groups, depending on their influence in the community. It is expected that the change of communities' structure should also affect community similarity. As the structure of communities proposed in this paper reflects the organization of individuals, without loss of generality, in the present study we refer such structure as the layered structure.

As the structure of communities categorizes nodes into different groups based on their influence, the shift of communities' structure can be measured by the shift of all nodes' influence. Thus, two kinds of weights are defined in our algorithm to measure the influence of nodes and the shift of the nodes' influence, respectively. The first one is the importance weight. It emphasizes that the core nodes could have stronger effect on the similarity of communities than on that of ordinary nodes. For instance, the overlapping of the core nodes could more strengthen the similarity of two journal clusters than that of the ordinary nodes with the same amount. The second one is the persistence weight. It proposes that the shift of nodes' status will weaken the similarity of two communities. For instance, if the core journals in one journal cluster degenerate to be ordinary in the next period, the similarity of the two journal clusters will be weakened.

We apply the novel algorithm to the journal data set from Web of Science Core Collection (WoS) database, aiming at tracking the changes in journal clusters and identifying anomalous events in science mapping.

The contribution of this paper is two-fold. First, we present a new community similarity algorithm, which takes not only the shift of the vertices, but also the shift of community structure into consideration. It is able to identify the transition of community structure in a better way, and is consistent to the intuition. Second, we apply the novel algorithm to a publication set. By calculating the similarity between each pair of two consecutive scientific data set, we are able to detect anomalous events in scientific mapping. Experiments show that there are two types of events in scientific mapping. The first one is related to a structural change in the journal clusters, while the vertex set remains relatively stable. It indicates that the research topics of the journal clusters may have changed. The second type reflects a situation, where the structure of journal clusters remains stable, though their vertex set considerably changes. This indicates that the new journals appearing in the set do not affect the structure of the journal clusters.
Related works

The evolution of scientific mapping

Literature about the evolution of scientific mapping includes, for instance, the evolutionary analysis of bioinformatics (Janssens, Zhang & De Moor, 2007; Janssens, Glanzel & De Moor, 2007), medical knowledge (Mina, Ramlogan & Tampubolon, 2007) and Health Research (Fiordelli, Diviani & Schulz, 2013).

Among these articles, we find three main approaches. The first category analyzes the lexical analysis of the scientific data set to detect emerging trends and bursts (Mane & Borner, 2004; Chen, 2006). Mane (2004) focused on identifying mapping topics and topic bursts in the Proceedings of the National Academy of Sciences (PNAS). They conducted Kleinberg's burst detection algorithm (Kleinberg, 2003), co-word occurrence analysis and graph layout techniques to generate maps that support the identification of major research topics and trends.

The second approach analyzes the evolution based on network theory using citation links. Boyack (2009) mapped the structure and evolution of chemistry research over a 30 year time frame. They analyzed the maps of 14 journal clusters in the field of chemistry research and visualized the knowledge flows among these journal clusters. They found that Biochemistry and Bioengineering are moving steadily into Chemistry territory, and are having a large influence on the general knowledge base.

Finally, a third approach makes an advantage of both methods by combining the two approaches. The so-called hybrid approach is a combination of citation-based and lexical analysis. This method was applied to track the evolution of research fields (Janssens, Glanzel & De Moor, 2007) and to detect and label emerging topics (Glanzel & Thijs, 2012).

In this paper, we conduct the citation analysis to depict the evolution of scientific mapping and detect scientific events. The major difference between our approach and the previous results is that we also consider the above-mentioned structural shift in journal clusters. The structure of journal clusters categorizes each journal of the cluster into distinct groups based on its influence. When the structure of a journal cluster changes, it indicates that a topic shift is likely to occur in the journal cluster. Thus, it is expected that the structural changes within journal clusters correspond to anomalous events in scientific mapping.

Graph similarity

Two types of graph similarity problems can be distinguished: (1) with known node correspondence; (2) with unknown node correspondence.

For the first type, various kinds of algorithms have already been proposed. Papadimitriou (2010) propose 5 similarity measurements for web graphs. Among them, the best two are the Signature Similarity, which is based on the SimHash algorithm, and the Vertex/Edge Overlap similarity. Bunke (2007) proposed techniques that were used to track sudden changes in communications networks for performance monitoring.

For the second category, the most widely used algorithms include: $\lambda$-distance (Bunke, Dickinson & Kraetzl, 2007; Peabody, 2002; Wilson & Zhu, 2008), which is based on the spectral method; algebraic connectivity method (Fiedler, 1973) and various algorithms based on graph kernels (Kang, Tong & Sun, 2012).

There are two major differences between the existing graph similarity algorithms and the novel approach in this paper. First, the existing algorithms usually ignore the shift of the layered structure of graphs. Second, communities do not necessarily possess sub-modularity structures, compared to graphs and networks. Therefore, the existing algorithms could not directly be applied to communities and a more specific method is required.
Methodology

In order to overcome the above-mentioned issue, we propose a new community similarity algorithm. We first present the basic idea of this community similarity algorithm. One key point in the algorithm is to measure the influence of nodes in communities, the shift of which is able to reflect the change in the communities’ layered structure. In this paper, the $k$-shell decomposition method is adopted to measure the nodes' influence. Then we introduce the algorithm itself.

The basic idea of the novel community similarity algorithm

The algorithm assumes that community similarity is affected by both the changes in their vertex set and changes in their structure. Intuitively, one assumes that communities are organized with certain structures, such as the flat structure and the hierarchical structure. The structure categorizes nodes into different classes, which reflect the influence of nodes. Thus, any change of a communities' structure can be measured by the changes of all nodes' influence.

For one node, $u$, belonging to two communities, $C_1$ and $C_2$, the following properties should hold in an effective community similarity algorithm:

1). If $u$ is influential in both $C_1$ and $C_2$, it will strengthen the similarity of $C_1$ and $C_2$.

2). If $u$ is influential in one community, but not so in another community, it will weaken the similarity of $C_1$ and $C_2$.

3). If $u$ is not influential in both communities, its influence on the community similarity is weak.

These properties show that community similarity is affected by two factors. The first one is the influence of nodes. Influential nodes are expected to be more crucial in community similarity measurement. For instance, when measuring the changes of one social group, the shift of group leaders should be more decisive. Thus, an importance weight is determined to measure the influence of nodes in the community. The second point is the change of the nodes' influence. The fact that a node is influential in one community but becomes an ordinary one in another community will weaken their similarity. Thus, a persistence weight will be defined to measure the change of a node's influence.

The Measurement of nodes' status

Both the importance weight and the persistence weight depend on the influence of nodes on communities. We adopt the $k$-shell decomposition method (Dorogovtsev, Goltsev & Mendes, 2006) to measure the influence of nodes. As proposed by Lu (2016), the $k$-shell decomposition method could identify the coreness of nodes, which is an effective indicator for the nodes' influence power.

The $k$-shell decomposition method is carried out as follows. First one removes all nodes with degree $k = 1$ from the network, and assigns the integer $k_s = 1$ to them. This procedure is repeated iteratively until only nodes with degree $k \geq 2$ are left in the network. Subsequently, one removes all nodes with degree $k = 2$ and assigns the integer $k_s = 2$ to them. Again, this procedure is repeated iteratively until only nodes with degree $k \geq 3$ are left in the network, and so on. This routine is applied until all nodes have been assigned to one of the $k$-shells.

Figure 1 shows an example of the community structure obtained by the $k$-shell decomposition method. In Figure 1, nodes in this community are categorized into three classes. Nodes with $k_s = 3$ are the most influential, and belong to the 'upper' class. These nodes are densely connected. Nodes with $k_s = 1$ are the least influential, and belong to the 'lower' class. They are usually located at the periphery of the community. The rest of the nodes belong to the 'middle' class.
Unfortunately, the $k$-shell decomposition method can only be applied to unweighted networks. To solve this problem, a modified $k$-shell decomposition method for weighted networks (Garas, Schweitzer & Havlin, 2012) is adopted. In particular, the weighted $k$-shell decomposition method applies the same pruning routine as the $k$-shell decomposition method above, but is based on an alternative measure for the node degree. This measure considers both the degree of a node and the weights of its links, and assigns a weighted degree, $k'$ to each node. The weighted degree of a node $i$ is defined as:

$$k'_i = [k_i^\alpha (\sum w_{ij})^\beta]^{\frac{1}{\alpha + \beta}}$$

where $k_i$ stands for the degree of node $i$, and $\sum w_{ij}$ stands for the sum of node $i$’s link weights. For simplicity, both $\alpha$ and $\beta$ are put 1, so that the weighted degree of a node $i$ can be written as:

$$k'_i = \sqrt{k_i \sum w_{ij}}$$

It is obvious that if the weights of all edges are 1 in the networks, the weighted degree reduces to the classical degree.

![Figure 1. The structure of a community obtained by $k$-shell decomposition method](image)

The structure based community similarity algorithm

We adopt the $k$-shell decomposition method to measure the influence of nodes. We conduct experiments on various networks and find that the number of nodes with certain $k$-shell value decreases exponentially as the $k$-shell value increases. Based on this fact, we assume that the importance weight of one node should also increase exponentially with its $k$-shell value. Thus, given the influence of the node $i$ in both communities, which are denoted as $k_{i1}$ and $k_{i2}$, the importance weight of node $i$ is defined as:

$$\text{weight}_i = 2^{(k_{i1} + k_{i2})/2 - 1}$$

which indicates that one node has large importance weight when it is influential in both two communities.

The persistence weight of one node is defined as:

$$\text{weight}_p = e^{-\tau(k_{i1} - k_{i2})/2}$$

where $\tau$ is the coefficient that determines the influence of the shift of a node's status. It shows that when the influence of one node remains unchanged in two communities, the persistence weight of the node is 1. The persistence weight of one node will considerably decrease, when the influence of the node in two communities largely differs.

For two communities $C_1$ and $C_2$, we denote $O(C_1, C_2) = C_1 \cap C_2$ and $F(C_1, C_2) = C_1 \cup C_2$. For each node $i$ in the set $O(C_1, C_2)$, the influence of the node in the two communities are denoted by $k_{i1}$ and $k_{i2}$, respectively. Thus, the overall weight of node $i$ is defined as:

$$2^{(k_{i1} + k_{i2})/2 - 1} \times e^{-\tau(k_{i1} - k_{i2})/2}$$
Thus, the overall similarity of two communities is defined as:

\[ \text{sim}(C_1, C_2) = \frac{\sum_{i \in \partial(C_1, C_2)} 2^{(k_{i1} + k_{i2})/2} - 1 \times e^{-\tau(k_{i1} - k_{i2})/2}}{\sum_{i \in \partial(C_1, C_2)} 2^{(k_{i1} + k_{i2})/2} - 1} \]

where the denominator is a normalization term, ensuring that the similarity of two communities takes values in the interval [0,1].

The publication data set

In this section, we briefly describe the journal set from the Clarivate Analytics Web of Science Core Collection (WoS) database, and how we use the journal data set to depict the evolution of subjects and detect the anomalous events in the cognitive mapping of science.

Data Sources and Data Processing

We apply the novel algorithm on the scientific data set from Clarivate Analytics. The original journal data contains tens of millions of so-called citable papers from 1992 to 2011 (i.e., papers of document type article, letter, note and review) indexed in the WoS database. We subdivide the dataset into four distinct periods, with each five years forming a single period. For each period, citations received by these papers have been determined for a variable citation window, which is consistent with the time interval of the period. By aggregating the citations among papers to the citations among journals, we obtain four journal networks, with nodes representing journals and edges representing the citations among journals. In particular, the number of both journals and citations in each journal network are given in Table 1.

### Table 1. The number of both journals and citations in the four periods of the dataset (Data sourced from Clarivate Analytics Web of Science Core Collection)

<table>
<thead>
<tr>
<th>Period</th>
<th>Number of journals</th>
<th>Number of citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992-1996</td>
<td>5434</td>
<td>1,273,292</td>
</tr>
<tr>
<td>1997-2001</td>
<td>6585</td>
<td>2,005,372</td>
</tr>
<tr>
<td>2002-2006</td>
<td>7164</td>
<td>2,834,616</td>
</tr>
<tr>
<td>2007-2011</td>
<td>8098</td>
<td>4,111,129</td>
</tr>
</tbody>
</table>

Reference Labels of Journals

To evaluate the effectiveness of the novel algorithm in detecting anomalous events in scientific mapping, we refer to the 23 categorizations of Essential Science Indicators (ESI). For each ESI field, we first extract both the set of journals that belong to the field and the cross-citations among these journals. Thus in each period, we obtain one community for each ESI category, and apply our new algorithm to monitor the evolution of each ESI category. In particular, for each ESI field, we obtain three similarity values in the four consecutive periods. By identifying the striking similarity values, we aim at identifying the anomalous events in the evolution of the 22 standard ESI fields.

Subject classification scheme by journal clustering

Most existing subject classification schemes, including the ESI categorizations, are based on longstanding practice (Zhang, Janssens & Liang, 2009) and are drawn up by human experts. Due to their purpose, which requires robustness and stability, these cognitive classification schemes usually remain unchanged for years. However, cognitive clustering of the document space can help improve subject classification by taking recent structural development into account (Zhang, Janssens & Liang, 2009). In this study, we follow this approach by developing and applying out new method.
Thus, we apply clustering algorithms to the four journal networks to obtain journal clusters, which could reflect subject classification and its changes over time. In particular, the spectral clustering algorithm (Von, 2007) is adopted. In this paper, we assume that the subject classification remains unchanged in a short period, and obtain 60 communities in each period. Similarly, we apply our new algorithm to these communities, and try to identify the anomalous events in these subject classifications.

Experiments on the ESI subject classification

In this section, we apply our algorithm to the 22 standard ESI fields in the four periods, aiming to detect the anomalous events indicated the two types of changes that we have discussed in the previous sections.

Figure 2. The distribution of community similarity for the journal clusters over time (Data sourced from Clarivate Analytics Web of Science Core Collection).


Figure 3. The word clouds for the 2nd ESI field (Data sourced from Clarivate Analytics Web of Science Core Collection)

(a) Period 1997-2001                  (b) Period 2002-2006                  (c) Period 2007-2011

Figure 4. The word clouds for the 20th ESI field (Data sourced from Clarivate Analytics Web of Science Core Collection)

We first apply the algorithm to obtain the community similarity values among consecutive communities for each ESI field. Then we determine the average similarity value and the
standard deviation of these similarity values. We denote the average similarity values by $\mu$ and the standard deviation of the similarity values by $\sigma$. Then the lower and upper limits of the normal similarity values correspond to $\mu \pm 3\sigma$. Using this method, we are able to define the threshold (lower control limit), below which the corresponding event is considered anomalous. In particular, the average similarity value is 0.86, the standard deviation is 0.0855, and the threshold amounts to 0.6035.

Figure 2 depicts the community similarities between consecutive communities for each ESI field. For the majority of subject fields, the community similarities lie in the interval [0.7,0.9]. Only three cases are under the threshold curve. The first case stands for the similarity between the first and the second periods for the 2nd ESI field. The remaining two cases stand for the two similarity values obtained in the last three periods for the 20th ESI field.

As shown in Figure 2, an anomalous event occurs in the 2nd period for this ESI field. To illustrate this, we present the word clouds of in both time slides. As shown in Figure 3, the 2nd ESI field in the 1st period focuses on the classical philosophy and history philosophy. Meanwhile, in the 2nd period, it focuses on different sub-disciplines of philosophy, including ethics, logic and phenomenology. Furthermore, Figure 3(a) indicates that America and Europe play a significant role in these topics of philosophy. By contrast, Figure 3(b) indicates that researches from Canada and UK in philosophy are more influential in philosophy in the second period. In turn, continental Europe’s influence on the research of philosophy has decreased in the 2nd period.

The 20th ESI category is concerned with ‘Plant & Animal Science’. As shown in Figure 2, anomalous events occur in the 3rd and 4th period. Thus, we present the word clouds of the three communities for this ESI field in Figure 4. Figure 4(a) and Figure 4(c) show that both communities focus on ‘entomology, biology, plant, physiology’. By contrast, Figure 4(b) shows that this community focuses more on ‘pathology, molecular, plantarum, and phytopathology’. This reveals that in the 3rd period researchers are more specialized on the research of treatment of plant disease. However, this specialization seems not to be persistent and disappears in the 4th period.

These two cases show that the new algorithm is efficient in identifying anomalous events in scientific mapping, and the identified events indicate topic transitions within each ESI field.

**Subject classification based on journal clustering: An experiment**

We adopt journal clustering techniques to obtain a subject classification scheme, and apply our new algorithm to analyze the evolution of the subject classification scheme in this section.

![Figure 5. Comparison of community similarities measured by both the new algorithm and the vertex overlapping strategy (Data sourced from Clarivate Analytics Web of Science Core Collection)](image-url)
Experimental results

We apply both the vertex overlapping strategy and the novel algorithm to calculate the similarities of journal clusters from adjacent periods. The comparison between the two strategies is plotted in Figure 5. The x-axis stands for the similarity of each pair of journal clusters obtained by the vertex overlapping strategy, the y-axis stands for the similarity of journal clusters obtained by our new algorithm, and the red line stands for the plot $y = x$.

Table 2. Text annotations of communities in Figure 5 (Data sourced from Clarivate Analytics Web of Science Core Collection)

<table>
<thead>
<tr>
<th>Case</th>
<th>Communities</th>
<th>Text annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A1/A2</td>
<td>philology, classical, historical, roman</td>
</tr>
<tr>
<td>B</td>
<td>B1/B2</td>
<td>biblical, catholic, jewish, fiction</td>
</tr>
<tr>
<td>C</td>
<td>C1</td>
<td>zoology, ecology, biology, evolutionary</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>entomology, zoology, insect, invertebrate</td>
</tr>
<tr>
<td>D</td>
<td>D1</td>
<td>information,library,documentation,electronic</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td>law, criminal, harvard, legal</td>
</tr>
</tbody>
</table>

Table 3. Core journals of communities in Figure 5 (Data sourced from Clarivate Analytics Web of Science Core Collection)

<table>
<thead>
<tr>
<th>Case</th>
<th>Communities</th>
<th>Top journals</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A1/A2</td>
<td>1.American Journal of Philology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.Arethusa</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.Classical Philology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.Classical Quarterly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.Journal of Roman Studies</td>
</tr>
<tr>
<td>B</td>
<td>B1/B2</td>
<td>1.Biblica</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.Catholic Biblical Quarterly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.Biblische Zeitschrift</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.Zeitschrift Fur Deutsches Altertum Und Deutsche Literatur</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.Zeitschrift Fur Die Alttetmentliche Wissenschaft</td>
</tr>
<tr>
<td>C</td>
<td>C1</td>
<td>1.Zoology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.Zoological Studies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.Evolutionary Ecology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.Animal Biology</td>
</tr>
<tr>
<td>C</td>
<td>C2</td>
<td>1.Entomological Science</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.Applied Entomology and Zoology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.Insect Systematics &amp; Evolution</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.Systematic Entomology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.Bulletin of Entomological Research</td>
</tr>
<tr>
<td>D</td>
<td>D1</td>
<td>1.Information Processing &amp; Management</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.Library &amp; Information Science Research</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.Journal of Information Science &amp; Evolution</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.Library Trends</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.Journal of Academic Librarianship</td>
</tr>
<tr>
<td>D</td>
<td>D2</td>
<td>1.Harvard Law Review</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.American Journal Of International Law</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.American Criminal Law Review</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.Law &amp; Society Review</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.Washington Law Review</td>
</tr>
</tbody>
</table>
Figure 5 shows that for most pairs of communities, the two similarity values are located around the plot $y = x$. This indicates that the similarities remain unchanged for the vertex overlapping strategy as well as our new algorithm. However, there are some exceptions. Four cases are marked, as their similarities obtained by the vertex overlapping strategy and our algorithm distinctly differ. The visualization, terms and core journals of communities for these four cases are presented in the following sections.

Case studies

For case ‘A’ and ‘B’, the structural similarities obtained by our new algorithm are much larger than that obtained by the vertex overlapping strategy. This indicates that the vertex sets in these communities considerably changes. By contrast, the structure of the community remains stable. In particular, for case ‘A’, the similarity obtained by our algorithm amounts to 0.60, while the similarity obtained by the vertex overlapping of 0.15 is much smaller than that obtained by our algorithm. For case ‘B’, the similarity obtained by our algorithm is 0.40, which is also much larger than that obtained by the vertex overlapping strategy.

We present both the terms and the most influential journals, particularly the core journals, of the two communities in Table 2 and Table 3, respectively. It can be seen that the core journals in these two communities are identical, and these are actually concerned with philology. One of the core journals is, for instance, ‘American Journal of Philology’, which was founded in 1880 and has achieved worldwide recognition as a forum for international exchange among classicists and philologists by publishing original research in classical literature, philology, linguistics, history, society, religion, philosophy, and cultural and material studies. In addition, the terms of both communities in case ‘A’ are ‘philology, classical, historical, roman’, which is consistent with the core journals. Furthermore, the word cloud figures for the two communities in case ‘A’ are also given in Figure 6 to provide more detailed information.
In case ‘B’, though the majority of nodes have changed, the set of core nodes remain stable. Thus, the similarity of the two communities based on the new algorithm is larger than the classical vertex overlapping strategy. Specifically, the two communities in case ‘B’ are concerned with ‘biblical, catholic, Jewish, fiction’. In addition, their core journals are identical, as well.

In cases ‘C’ and ‘D’, the similarities obtained by the vertex overlapping are much larger than that obtained by the novel algorithm. This indicates that although the vertex set of the communities remain relatively stable, the internal structure of these communities changes significantly. A potential topic shift has occurred in these communities.

As shown in Figure 5, the two communities in case ‘C’ share 35% of their vertices, while their structural similarity is 0.14, which is much smaller than that obtained by the vertex overlapping strategy. This indicates that the layered structure of these two communities considerably differs. According to Table 2, though the two communities in case ‘C’ concerned with zoology, they focus on different sub-disciplines. The first community could be characterized as being focused on ecology and biology, while the second community is concerned with animals, such as entomology, insects, and invertebrate. Furthermore, the core journals of the two communities also differ, as shown in Table 3. In addition, the word cloud figures for the two communities in case ‘C’ are given in Figure 7.

Similarly, in case ‘D’, the 1st community focuses on the library and information science. However, the 2nd community focuses on law and criminology. The core journals of the two communities in Table 3 are different, which deepens this conclusion.

Conclusions

In this paper, a novel community similarity algorithm is proposed to detect structural changes and topic shifts in science mapping. This concerns changes of vertices and the communities’ layered structure. This makes it possible to detect communities with vertices remaining stable but structure changed. These kinds of events in scientific mapping indicate that a topic shift took place in the corresponding research fields (Glanzel & Thijs, 2012).

We applied our new algorithm to the journal data set. Our experiments indicate that the novel algorithm is able to identify meaningful changes in scientific mapping. We found that the internal structure of the majority of journal clusters remained stable over time. This indicates that most of the research subjects keep their research topics unchanged. Beyond that fundamental observation, we could identify several journal clusters with changes in their journal constitution but with stable internal structure. This indicates that although journals focusing on the corresponding research subjects considerably change, the topic of the journal cluster remains unchanged. We also found several journal clusters stable members but with changes in their internal structure. This indicates that a topic shift took place in the corresponding research subjects. The textual analysis of these journal clusters demonstrates that the identified events are meaningful.

The aim of this study was to introduce a new method for topic identification and monitoring and to show its validity and usefulness. In later studies, we will apply our new algorithm to analyze the evolution of concrete research fields. The identification of topic shifts and structural changes has the potential to improve techniques for the monitoring of evolutionary patterns of research fields and the detection of emerging topics.

References


Networks of international collaboration and mobility: a comparative study

Zaida Chinchilla-Rodríguez¹, Lili Miao², Dakota Murray², Nicolás Robinson-García³, Rodrigo Costas⁴, and Cassidy R. Sugimoto², ⁴

¹ zaida.chinchilla@csic.es
Consejo Superior de Investigaciones Científicas (CSIC) (IPP), SCImago Group (Spain)

² lilymiao08@gmail.com; dakota.s.murray@gmail.com; sugimoto@indiana.edu
Indiana University Bloomington (USA)

³ elrobinster@gmail.com
Universitat Politècnica de València (INGENIO-CSIC-UPV), Valencia (Spain)

⁴ rcostas@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (Netherlands)

Abstract
This study presents a preliminary comparison of networks of international collaboration and mobility. Using affiliation data from scientific publications, we analyse the structural differences in the two networks and the role of countries. The results show that researchers collaborate internationally to a much higher degree than they become internationally mobile. The number of countries involved in the networks is three times higher in collaboration than in mobility, and the average degree demonstrates that mobility networks form tight structures with fewer links than collaboration networks. The role of countries differs between the collaboration and mobility network, predominately reflecting income level. Limitations and future research are described to further understand the dynamics of collaboration and mobility networks.

Conference Topic
Scholarly Communication; scientific mobility; collaboration; network analysis

Introduction
There is a popular saying: “Science knows no borders”. Indeed, scientists often move and collaborate internationally, accessing “multiple learning environments”, usually producing higher-impact research, and increasing their prestige (Gazni, et al. 2012; Glänzel 2001; Chinchilla et al. 2012). There are many well-documented reasons why researchers collaborate (e.g., Beaver 2001; Sonnenwald, 2007). Previous studies focus on the factors driving collaboration between countries, such as their relative size, their geographical, historical, linguistic, and thematic proximity, or their socio-economic characteristics (Zitt et al. 2000; Adams 2014).

Mobility has been advocated as key to increasing the efficiency and effectiveness of research (OECD 2008, 2010; Scellato et al., 2015). While some studies focused on the economic and development impact caused by mobility (Gibson & McKenzie 2012), few have utilized bibliometric measures to understand scientific mobility (Moed & Halevi, 2014; Sugimoto, Robinson-Garcia, & Costas, 2016). Furthermore, no study to-date has compared, at a global scale, the differences in the scientific networks created through collaboration and mobility. This is necessary to understand whether mobility provides a novel or a duplicative lens on the network created through scientific collaboration.

We present here, using collaboration and mobility data drawn from contemporary bibliometric data (2008-2015), the comparative network statistics and ranking for more than 200 countries considering socioeconomic indicators and grouping according to scientific and technological
capacities. These data can inform the understanding of and relationship between indicators of international collaboration and mobility.

Data and Methods
Data were retrieved from Web of Science database for the period 2008-2015, considering only authors that have at least two publications and for whom this period represents the date of first publication. A total of 3,521,797 individual’s authors were identified using a large-scale disambiguation algorithm developed by Caron and van Eck (2014). We consider all document types for the analysis. The nation-to-nation links were collected based on 14,097,939 publications. Country affiliations were extracted and cleaned, resulting in 213 unique countries. Collaboration linkages are measured considering all authors of each paper and their connections to countries. Mobility includes both migration (i.e., change from one affiliation to another) and co-affiliation (having multiple affiliations on the same publication or on two or more publications in a given year). Income level group was added as a proxy of wealth intensity of countries (World Bank 2016) and the Scientific and Technological Capacity Index—which measures expenditure in R&D, number of researchers, number of institutions, patents--(Wagner et al., 2001) was used as a proxy for the scientific capacities of countries.

For each group, we created and analyzed the subset of documents and researchers by calculating: the total number of publications per country and the number of publications with international collaboration. We use these two indicators as input to the international collaboration network. We build the international mobility network using the total number of active researchers per country and the number of active researchers linked to countries. We used integer counting, attributing a count of “1” to each occurrence of authorship or affiliation from a country.

Once we obtained the co-occurrence frequencies, we generated symmetric matrices using Pajek 64 4.10 (Batagelj & Mrvar, 1997) to derive statistical properties and structure of the networks. We explore the position and role of countries within each group using classical measures of network analysis. Closeness is calculated as the reciprocal of the sum of the shortest path between the nodes to other nodes (Freeman, 1978). Betweenness for a node is defined as the fraction of shortest paths among the whole network that pass through the node (Freeman, 1977). Density determines the degree of cohesion that exists among the nodes, revealing whether the network has a thick or thin consistency (Wasserman & Faust, 1999). The average degree measures the spread of influence across the networks (Hanneman & Riddle, 2006). Diameter measures the longest distance between a pair of countries, (i.e., how many steps from any node are necessary to reach any other node in the network (De Nooy et al., 2011)). The clustering coefficient indicates the proportion between the number of links in the neighborhood of a node and the number of links possible in the entire network (Watts & Strogatz, 1998; Barábasi, 2002). Assortativity is a property that reflects the tendency of nodes to connect to other nodes to a similar degree by means of measuring the Pearson correlation coefficient of degree between pairs of linked nodes (Newman, 2002).

For the visualization network, the proportions of collaborations and co-affiliations among countries were used as a pair of inclusion indexes (share of county j in the total co-authorships/co-affiliations of i or Affinity Index AFI (i,j), and its counterpart AFI (j,i) to characterize asymmetrical relationships between two countries (Zitt et al., 2000). AFI is a measure of the amount of collaboration or co-affiliations between a given country A and another country B, compared to the total collaboration or co-affiliation of the country A and country B. To position the countries (nodes) we applied the Kamada-Kawai algorithm (1989),
characterized by its ability to assign coordinates to the vertexes while adjusting the distances to a maximum with respect to theoretical distances. Lines indicate the relations among actors, with the color (on a scale of grey tones) indicating the intensity of the connection. The colors of countries refer to geographical regions.

Results
Each network, summarized in Table 1, has 212 countries, one less than the original 213 because South Sudan is isolated in the collaboration network and Tuvalu is isolated in mobility. Differences in the number of edges, the density, and the average degree of each network suggests significant differences in their structural cohesion. The collaboration network has a higher density than the mobility network. Nodes in the collaboration network have three times the average degree than nodes in the mobility network. In comparison, the average shortest distance varies only slightly, indicating that the two networks are cohesively connected, where countries could reach each other in an average of less than two steps. However, considering their diameter, a distinctive feature appears: the collaboration network has a diameter of two, meaning that it is tightly linked with short distances between any pair of nodes; however, in the mobility network the diameter is twice as large, suggesting that the two networks are structurally different. The clustering coefficient is high in both networks demonstrating that countries form tightly knit groups, but the coefficient is slightly higher in collaboration than in mobility, indicating that two countries who have common neighbors would be slightly more likely to collaborate than have mobile researchers in common. However, considering that there are more than 2.5 times as many links in collaboration as in the mobility network; the mobility network has apparently formed tighter structures with fewer links. The negative values of assortativity indicate that countries with a large degree tend to connect with countries with a small degree in both networks.

<table>
<thead>
<tr>
<th>Table 1. A summary of the properties of the two networks.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Collaboration</strong></td>
</tr>
<tr>
<td>Number of nodes</td>
</tr>
<tr>
<td>Number of edges</td>
</tr>
<tr>
<td>Density</td>
</tr>
<tr>
<td>Average degree</td>
</tr>
<tr>
<td>Average shortest distance</td>
</tr>
<tr>
<td>Diameter</td>
</tr>
<tr>
<td>Cluster Coefficient</td>
</tr>
<tr>
<td>Assortativity</td>
</tr>
</tbody>
</table>

The rank of countries in collaboration and mobility networks
Figures 1-4 show the position of countries in closeness (left), betweenness (center) and clustering coefficient (right) for each network: with collaboration in the left panel and mobility in the right. We show four groups of countries according to their scientific and technological capacities where the colors of lines represent the income levels of countries: high (green), upper-middle (blue), lower middle (orange) and low-income (red). The interpretation of ranking labels goes as follows: Australia (3) (5) means that it ranks in the third position in the collaboration network and in the fifth position in the mobility network.
In the group of scientifically advanced countries (Figure 1), a country’s position is relatively stable across the two networks, suggesting that there exists a relatively similar ability to attract collaboration and mobility partners. The most striking situation is in Russia (upper-middle income) ranked relatively low in closeness and betweenness, whereas it has the second highest clustering coefficient in the mobility network; this suggests that Russia is not located in a central position, instead forming a tight group of partners, especially in mobility. While France, Germany, Denmark, Finland, and Ireland have more capacities to reach other countries and act as bridges especially in the mobility network, Canada, Taiwan, and Iceland show a more important role in the collaboration network.

Figure 2. Closeness, betweenness and clustering coefficient for scientifically proficient countries
The positions of scientifically proficient countries fluctuate more than those of advanced countries (Figure 2). India ranks higher in closeness and betweenness in the mobility network than in the collaboration network, but their group of partners in collaboration is more cohesive than in mobility. That means that the country acts as a key bridge in the mobility network, but its mobility neighbors do not form tightly group around it. A similar situation appears in Slovenia and Croatia, whereas Estonia, Bulgaria, and Cuba have roles that are more important in international collaboration than in mobility.

In the group of developing countries (Figure 3), Mexico and Chile rank highly in closeness and betweenness, but their slight drop between collaboration and mobility demonstrates that they can build and facilitate more connections in collaboration than in mobility. Turkey shows the opposite pattern, being more important in mobility. These three countries show the lowest cohesion with their partners in comparison with the rest of the group. Iran’s, Pakistan’s, and Armenia’s closeness and betweenness rank in mobility suggests than they have most important roles in linking researchers through mobility than collaboration. A striking case is Turkmenistan, which ranks at a top position by cluster coefficient and in the lowest position in closeness and betweenness, suggesting that while it does not act as a bridge in the network, it does forms a tight collaboration relationship with its partner compared to other countries in its group. Overall, developing countries are less stable between networks that advanced and proficient countries.

Countries classified as lagging are the most numerous (80 countries) and for visualization, we only show the top 30 in each indicator, explaining the lack of correspondence among the appearance of countries (Figure 4). The only high-income country appearing in this sample (among the six marked as lagging) is Saudi Arabia, which has an important role in closeness and betweenness, especially in the mobility network, whereas Malaysia leads in these two indicators in both networks. However, neither Malaysia nor Saudi Arabia establishes a tight
relationship with their partners in comparison with the rest of lagging countries. In general, many countries’ position changes dramatically, indicating that their positions and roles differ between the networks.

Figure 4. Closeness, betweenness and clustering coefficient for scientifically lagging countries

Both the collaboration and mobility networks demonstrate the importance of geographical proximity in the topology of the network (Figure 6). Historical and political linkages are also apparent—for example, the links between countries and colonies as well as linkages in former Soviet countries. Language also matters, as demonstrated by the connections between Spain and Latin American countries. The two networks the most advanced countries occupy the center, but the change in the position of some Asiatic countries such as India, Taiwan, Singapore, Malaysia or the position of South Africa, Nigeria, Kenya in the two networks reveal significant differences in partnership preferences but also in policies that promote the intensification and expansion of international scientific networks.
Figure 5. Network of international collaboration
Figure 6. Network of mobility
Discussion and Conclusion
This study presents a preliminary comparison of the diversity and complexity in establishing international relationships in terms of collaborative papers and the flow of mobile researchers. The results show that researchers establish more international relationships through collaboration than mobility. To a certain extent, networks show a high level of local clustering and a small average number of steps between actors, which fits with the model of the small world defined by Watts and Strogatz (1998) and the results of Wagner (2005).

However, the mobility network forms tighter structures with fewer links than the collaboration network, making this a distinctive feature between the two networks, despite the diameter of collaboration network is tighter than in mobility network. A country’s position in collaboration and mobility is not always consistent in both networks, revealing that positions and roles of countries in terms of capacities to be accessible to other nodes (closeness), in transferring knowledge (betweenness), or to be cohesive with their neighbours (clustering) may differ. Where stability between the networks exists, it is mostly associated with high-income countries, while upper-middle, lower middle, and low-income countries are more likely to change positions. Unlike countries in the scientifically advanced group that function as hub nodes, the role of lagging countries is unstable in both networks. However, the stability of a small subset of countries as bridge nodes, such as India, Malaysia, Turkey and South Korea, builds stronger connections in the mobility network. We also confirm the important role of Kenya and Nigeria, as noted in previous studies relating to collaboration networks (Adams et al., 2014). Although leading research economies tend to attract more researchers in terms of collaborative papers and mobility, the results suggest that there are complex patterns of knowledge circulation. The representation of the two networks also displays interesting drivers that reveal affinities between different economies based on linguistic, historical, as well as political and cultural linkages.

Further analysis is needed to overcome some of the limitations of this study and respond to other important questions related to the capacities and influences of countries in networking science. For example, the analysis of co-affiliations in Web of Science has an inflationary effect on the traditional measurements of collaboration based on affiliations creating some overlaps when we are comparing collaboration and mobility. There is a major general limitation of collaboration analysis based on author affiliations and further analysis should be done to minimize this effect (Hottenrott and Lawson, 2017). We intend to complement our analysis with a time component, allowing us to analyze some key points: authors’ choices regarding institution address selection from publication data and the evolving relationships for exploring how these more fine-grained temporal networks compare to the entire network presented here. We will explore different approaches to community detection in networks to identify the core and groups of vertices having the highest probability of holding a great deal of influence over the organization of the periphery of the network. We plan to analyze not just mobility, but also leadership, including changes in the positions that authors occupy in the bylines of co-authorship, their impact, the institutional reputation of destinations, and the capacity to develop or reinforce thematic research into their institutions and countries. In addition, the analysis of other factors that would influence the mobility and collaboration of researchers such as cultural, linguistic and geographical proximities will be analyzed. Overall, this analysis presents an introduction to what is possible through comparison of international collaboration and mobility networks, and serves as a foundation for future analysis.
Acknowledgments

Financial support from Mobility Program ‘Salvador de Madariaga 2016’ and State Programme of Research, Development and Innovation oriented to the Challenges of the Society (Ref. CSO2014-57770-R) funded by the Ministry of Economy and Competitiveness of Spain and the Science of Science Innovation and Policy program of the National Science Foundation in the United States (NSF #1561299).

References


Study on Ubiquitous Network Intelligent Service Based on Situation Fusion KDD

Liu Yong¹ Ren Yan²

¹ y_liu@zzia.edu.cn
Zhengzhou University of Aeronautics, Zhengzhou (China)
Collaborative Innovation Center for Aviation Economy Development, Zhengzhou (China)

² renyan421@126.com
Zhengzhou University of Aeronautics, Zhengzhou (China)
Collaborative Innovation Center for Aviation Economy Development, Zhengzhou (China)

Abstract
Intelligent service is the inevitable trend of ubiquitous network development. Situation Fusion Knowledge Discovery in Database(KDD) technology can further provides the intelligent service of ubiquitous network. In this paper, based on the characteristics of general structure of ubiquitous network, the method of Situation Fusion KDD technology is proposed for the problem of ubiquitous network information demand and drift of network element supply. By improving the information description and mining algorithm, the process of Situation Fusion KDD technology, such as situation role acquisition, situation role behaviour analysis, front-end application information mining and situation requirement knowledge discovery, is discussed in this paper. According to the functions which are involved in the ubiquitous network intelligence service of situation analysis including data organization, pattern discovery, resource aggregation and intelligent push, the application method of the Situation Fusion KDD is analyzed and the realization path of the intelligent service in the ubiquitous network is provided finally in this paper.

Keywords
Situation fusion; Knowledge discovery in database; Ubiquitous network; Intelligent service

Conference Topic
Knowledge discovery and data mining

Introduction
The information and demand are ubiquitous in the ubiquitous network where the information carrier forms, data sources, data structures, information content, quality standards, data exchange formats, etc. have become very complex. How to provide timely and accurate information and knowledge facing the network users’ complex demand with dramatic changes of roles is the main problem of ubiquitous network to solve. This paper argues that Situation Fusion Knowledge Discovery in Database(KDD) technology is the key to solving this problem. In this paper, according to the characteristics of user demand in the ubiquitous network, the KDD technology was used to study the principle of Situation Fusion KDD and explore the innovation mode of ubiquitous network intelligent service, thus solving the problem of information acquisition precision and service performance in the ubiquitous network.

Issues on Ubiquitous Network Intelligent Service
Ubiquitous network is the whole space-time interconnecting network based on the application support services as the main feature. Compare with the traditional information network, the same point are their four major constituent elements: human, material, capital and information. Their difference is that the ubiquitous network is the new network phenomenon where the above four elements deeply integrate to a certain stage, which is the consequence due to the highly developed traditional
information network in the application and service level. In this way, the role demand, the follow-up of application and service, the realization of value and data flow together constitute a new form of space-time ubiquitous network with “point to point” barrier-free real-time interactive experience.

In the ubiquitous network, there are two kinds of changes in the four elements of human, material, capital and information. One is the changes of the elements themselves, namely, the changes of “human” in network roles, the changes of “material” in the intelligent perception, application and service forms, the changes of “capital” in value realization dimensions and the changes of “information” in the data transmission access types. The other is the changes between elements, that is, the four elements have the phenomenon of the conversion of main body and object and the phenomenon of supply drift.

In the ubiquitous network, the benefits of intelligent service are the free conversion of the main body and object, the more extensive source of information, the surge in the amount of data, high value of information correlation and the value-added effects of information processing. The new form of space-time ubiquitous network has also led to the migration from the individualized customization needs to the various roles’ situation needs as well as the phenomenon of the network elements supply drift. The traditional intelligent service technology and method can solve the information demand problems with the small change of the network elements, but it is difficult to play the role of the transformation among network roles and the elements supply drift, which is a new topic in the ubiquitous network intelligent service. Due to the relative stability of the situation in the ubiquitous network, it is very important to explore the application of principle and method of Situation Fusion KDD in intelligent service mode.

**Situation Fusion Knowledge Discovery in Database**

According to the definition of the International Telecommunication Union, the ubiquitous network is divided into five layers: the sensing layer, the access layer, the next generation network layer, the middleware layer and the application layer. Among them, the next generation network layer, middleware layer and application layer are the key points of technology research and application support in the ubiquitous network. The next generation network layer provides ubiquitous communication; the middleware layer provides ubiquitous resource integration; the application layer provides ubiquitous situation services.

It can be found that the ubiquitous network architecture has been reflected the design idea from the user-oriented service mode to the situation-oriented service mode.

It is clear that traditional user fusion knowledge discovery method is difficult to adapt to the new paradigm and new requirements of the ubiquitous intelligent service. To this end, the further study for the situation demand is needed to enhance the function and application of Situation Fusion KDD based on the traditional KDD techniques and methods.

**Principle of Situation Fusion KDD**

The main function of traditional KDD is to extract information and knowledge and provide decision support service from massive data for users’ individual needs. In order to solve the problem of transitional drift of information demand caused by excessive conversion of user roles in the ubiquitous network, it needs to be improved in the user-oriented aspect and the new mode, Situation Fusion KDD, is thus needed.

The basic principle of Situation Fusion KDD is to change the data description way from the traditional way of describing the users as objects into describing the situations as objects so as to improve the mode and algorithm of KDD to achieve the intelligent service function for the situation and meet the information and knowledge needs of situation users in order to provide decision-making advisory services.
Functions of Situation Fusion KDD
The purpose of Situation Fusion KDD is to discover implied and valuable knowledge from situation information and predict the future trends and behaviour of situation users, thus providing advanced, situation-based decision support. Situation Fusion KDD should have five functions, including role behaviour prediction, situation correlation analysis, situation role clustering, situation resource classification, sequence pattern discovery and so on.

Predict the behaviour trend of situation groups
Predict various situation user groups that are interested in all aspects of information in the future and predict their behaviour patterns.

Situation correlation analysis
Automatically classify the seemingly irrelevant but frequently accessed data according to different situation to find potential associations between situation users.

Situation role clustering
The relevant information is recommended to the situation group based on the user group characteristics of different situations or the data accessed by the situation user.

Situation classification
Find the situation information and situation access mode which can access to the front-end application to obtain the situation profile features. Classify situations by those features and form the situation information resource system as the basic data using Situation Fusion KDD technology to predict potential situation roles and provide situation information service.

Situation sequence pattern discovery
Forecast the user situation access mode and carry out intelligent information services facing the situation to ensure the stable access of the front-end situation application information and knowledge.

Process of Situation Fusion KDD

Get situation access information
Ubiquitous network situation information is divided into website-side access information and mobile-side access information. At present, as the rapid spread of mobile Internet and the surge in the number of mobile users, the feature of “ubiquitous” is reflecting more and more significantly in the mobile Internet. Therefore, the situation information collection should be carried out from both the website side and the mobile side. Crawl the front-end application information through the ubiquitous network agent and use the knowledge base filtering technology to filter out the disturbing or low correlation front-end situation application information and data. Then extract the theme, keywords, paragraphs, special logos and other important front-end application data, thus gaining valuable information that is useful for the analysis of situation role access behaviour.

Analyze behaviour of situation roles
Generate situation role multi-dimensional tags and attribute characteristics through behaviour analysis tools. Situation role multi-dimensional tags represent different situation types and related characteristics accessed by situation roles; situation role attribute characteristics are
established based on the situation role multi-dimensional tags and deeply excavated according to the different situation probability level. Then, the situation role attribute characteristics are obtained and should be discovered for clustering so as to generate the situation role groups. On this basis, the construction of the initial situation needs pattern is completed.

Dig out front-end application information
Combined with situation information, use the analysis engine tool to analyze front-end application content for processes such as text-cutting, clustering, knowledge processing.

Situation demand knowledge discovery
According to the situation demand, dig out the information and knowledge of the role users in the situation area to complete the dynamic updating of the situation knowledge base and the dynamic updating of the situation mode.

Ubiquitous Network Intelligent Service
Ubiquitous network intelligent service use Situation Fusion KDD technology to study situation behaviour and situation definition taking the situation roles as objects for offering high-quality intelligent service to the situation users. According to the situations of various carriers themselves, ubiquitous network intelligent service system not only has the function of providing intelligent service for situation users, but has self-learning, self-adapting and self-organizing functions to meet the information and knowledge needs of situation roles users.

Advantages of Ubiquitous network intelligent service
The traditional intelligent information service system can meet the different depth of the needs of different levels of users, but there exist shortcomings such as single channel and single access information because of the single user domain and professional and other limits. Ubiquitous network intelligent service use Situation Fusion KDD methods to more deeply discover the potential situation needs in a relatively stable situation type, providing multi-faceted situation information and situation knowledge services. At the same time, regardless of the drift of user individual needs and the changes of user individual roles, the service quality and efficiency will be higher only pushing specific information and services for the situation needs. Therefore, comparing with the intelligent information service system which takes the network users as objects, the network intelligent service which takes the situation roles as objects has more depth in level, more extensive breadth, more accurate quality, higher service, more viscosity in media.

Features of Ubiquitous Network Intelligent Service
Ubiquitous network intelligent service system uses the data mining, artificial intelligence and other technical means to carry out situation fusion knowledge mining and intelligent service based on the characteristics of situation roles and information of situation requirements in the ubiquitous network ecosystem, with the following six characteristics.

Cross-border excavation
According to the same situation, the system aggregates cross-domain data and analyzes data and then discovers deeper level psychological needs of situation roles through collaborative processing.
**Multidimensional fusion**
In the dimensions including demand dimension, information dimension, method dimension, carrier dimension and role dimension, the data fusion should be completed based on the same situation to form the stable data resource which is more conducive to the stable application of knowledge discovery tools.

**Cobweb interactivity**
The intelligent service system and the dialogue of situation roles have all-round ubiquitous characteristic. In the same situation role, whether managers, consultants or network users can all engage in dialogue and interaction with the system in the form of full media and spider response regardless of time, space and fixed identity. The system automatically collects information and accurately presents to the user to meet the needs of different situation role users for the users in different situations.

**Perceived enhancement**
The system uses a variety of situation-aware technologies and augmented reality technology to provide visual information services.

**Dynamic experience**
Situation information needs positioning, situation information organization, knowledge mining, information transmission, service space, service form and user experience and other forms are open, multi-dimensional and real-time dynamic.

**Predictive accuracy**
Based on the Situation Fusion KDD service system, due to the deep excavation of stable situations, learning and service, the situation definition analysis and identification prediction are more accurate and the intelligent services are more active.

**Ways of Ubiquitous Network Intelligent Service**

**Situation perception and situation data acquisition**
The collection of the situation role characteristics and situation demand information in the ubiquitous network can be carried out through situation perception and collaborative acquisition. The situation supply and demand information source is formed through intelligent access, intelligent analysis and information aggregation.

**Situation data organization**
Integrate the situation needs and role characteristics to build situation data resource self-organization mechanism and form the situation supply and demand resource pool. At the same time, integrate information resources and form dynamically updated data resources and knowledge resource system to complete the intelligent service data collection.

**Situation needs and service pattern discovery**
The main function of the situation demand service mode use the Situation Fusion KDD technique to analyze the characteristics of the situation roles, and to find the potential knowledge rules and the situation demand pattern.
By using the correlation analysis and sequence pattern analysis technique, the correlation and regression characteristics of situation supply and demand information are analyzed and the knowledge requirement rule and situation service mode are formed.
Situation resource and technology aggregation

Combined with artificial intelligence, DSS and ES, use Agent theory, heuristic and other intelligent decision-making algorithms to integrate situation data, situation modes, reasoning technology and other resources to form human-computer interaction ubiquitous intelligent service environment for the base of intelligent situation resources of knowledge decision service.

Ubiquitous network intelligent service

The system analyzes and processes the collected information by analyzing the user situation roles and mining the information of the situation role information demand and then provides the service for different situation roles according to different situation, transferring from user information transmission to situation knowledge service. Ubiquitous network intelligent service mainly includes intelligent retrieval service, intelligent recommendation service and intelligent consulting service. The intelligent retrieval service intelligently sorts out the information and knowledge in the unified situation category and initiatively provides the intelligent retrieval service oriented to the situation roles. Intelligent recommendation service is the highlight of the system, the use of situation requirements and role characteristics data helps to initiatively recommend the situation needs information and knowledge and strengthen the situation users’ experience. Intelligent consulting service offers other users the consultation of the same situation based on the same service case of user information and knowledge needs in the situation domain, thus, the matched problem is more accurate and service efficiency and quality are higher than before.

Concluding Remarks

In short, the ubiquitous network intelligent service is oriented towards situation needs. The traditional service mode is user-oriented. Although it also has intelligent retrieval service, intelligent recommendation service and intelligent consulting service and other service forms, it cannot solve the individualized service’s poor quality caused by the drift of user demand and network element supply in the ubiquitous network environment. Ubiquitous network intelligent service changes the user-oriented service into the situation-oriented service and changes the individualized service into the role-oriented services, highlighting the situation of the service to adapt to the new requirements of intelligent services in the ubiquitous network, thus, improving the efficiency and quality of intelligent services. Ubiquitous network intelligent service takes situation requirements, role characteristics, knowledge resources, network facilities and application platform as the organic information resource system to organize and process to separate the users from the resource system and more highlight the analysis and service of the situation roles, meeting the different needs of information and knowledge due to the users’ changing roles. This paper analyzes the principle of Situation Fusion KDD and applies the technology to the construction of ubiquitous network intelligent service system. It is a preliminary attempt to generalize the ubiquitous network intelligent service theory and implementation method in order to get more in-depth research and discussion.

References

Lian Jie. Research on User Features Based Data Mining in Social Networks[D]. Beijing Jiaotong University, 2014.

Fund Project
the National Social Science Foundation of China, No. 14BTQ070

Author information
First Author: LIU Yong, Ph D, professor, specializes in information visualization. E-mail: y_liu@zzia.edu.cn
*Corresponding author: REN Yan, Master degree candidate. E-mail: renyan421@126.com
How large is large enough?

Zhesi Shen¹, Liying Yang², Zengru Di¹ & Jinshan Wu¹

¹School of Systems Science, Beijing Normal University, Beijing, 100875, P. R. China
²National Science Library, Chinese Academy of Sciences, Beijing 100190, P. R. China

Abstract
It is well known that the Journal Impact Factor (JIF), originally intended as an indicator of impact of individual journals, should not be used as an unconditional standard of evaluation for individual researchers and individual research papers since the amount of citations of papers within any given journal has a highly skewed distribution such that the average number of citations are no longer typical. In this work, for each given journal, we ask the question that if one paper is not typical then a group of how many papers together can be typical. Typical means that the group average is close to the JIF while the standard deviation of average of the group generated by a Bootstrap process is much smaller than the standard deviation of single papers. According to this size, we proposed a definition of the minimum representative size of each journal. In addition, we found that this minimum representative size of most journals is significantly less than the corresponding journal size while for a few exceptions, this becomes comparable to the journal size or even bigger.

Conference Topic
Impact Factor, Journal-level indicators, Research Evaluation

Introduction
Evaluating the performance of scientists is important in promotion and grant applications, and a considerable amount of metrics have been proposed based on various assumptions (Bar-Ilan, 2008, Bornmann et al., 2012, Mingers & Leydesdorff, 2015, Waltman, 2016). As the most easily accessible metric, journal-level metrics (Glänzel & Moed, 2002), and more specifically, the Journal Impact Factor (JIF) (Garfield, 1955, 1999), of a journal in which their paper is published is used in the routine evaluation of individual scientists (or their work). The effectiveness of JIF in evaluating individual scientists and individual papers has been highly criticized in recent years (DORA, 2013, Hicks et al., 2015), with its main weakness being using an arithmetic mean to represent a highly skewed citation distribution (Seglen, 1992, 1997, Kurmis, 2003, Leydesdorff & Opthof, 2010, Factor, 2011, Bornmann & Mutz, 2011, Leydesdorff & Bornmann, 2011a,b, Mutz & Daniel, 2012) that highly overlaps among most journals (Larivière et al., 2016, Callaway, 2016, Mingers & Yang, 2017). Some journals have decided to not provide JIF on their journal webpage, or have chosen to provide a broader range of metrics (Callaway, 2016). We should note that, while metrics may facilitate performance of such an evaluation, peer review and in-person discussion remains the most reliable although much less feasible, method of evaluating scientists. With that being said, metrics do provide some convenience in comparing a set of papers or scientists in terms of, for example, journals or institutes. More specifically, it has been shown that, on average, journals with a higher JIF generally have higher impact papers than those with lower JIFs (Milojević et al., 2016). Of course, in much of this line of work, especially (Milojević et al., 2016), it is assumed that the number of received citation of a paper is a good proxy for the impact of the paper and we also take this assumption.

The principal reason JIF is not suitable for application to individual papers is the frequently highly skewed distribution of the amount of received citations for papers of each journal and to represent a highly skewed distribution function it requires moments at all orders, not only the mean, which is the first order moment. However, providing each journal such a set of
indicators of all orders of moments will then become overly complex. Therefore, one question that should be addressed is how to define a less complex indicator for journals that yields practical application not only to journals, but also to individual papers. There are such indicators proposed in the literature, such as the Integrated Impact Indicator and Percentile Rank (Bornmann et al., 2013, NSB, 2014).

Here, we decided to take an alternative approach to this question. From statistics (Wasserman, 2004), we know that given a set of data, we might calculate a series of K-sample averages μ(K) from the data. It means to take K values from the data with putting back and then to calculate the average of these K values. For example, the usual mean corresponds to 1-sample average. We also know that while μ(K) stays the same, μ(K) = μ (1), often the standard deviation Δ(K) decreases with K (Wasserman, 2004). For instance, if the data set follows a normal, distribution then Δ (K) ~ Δ(1)/K. This suggests that instead of looking at the reliability of using JIF on individual papers of the journal, we may look into how to reliably use JIF on a set of papers from the journal. If citation numbers of papers from a journal fluctuate extremely largely such that Δ(K) and higher order moments are very huge, then naturally, it will require a very large K to make Δ(K) to be sufficiently small. It is even possible that such K is required to be larger than the number of papers of the journal P, K > P. On the other hand, if papers from a journal have more or less the same number of received citations, then such required value of K might be very small. Thus, by looking at how large K is required to make Δ(K) sufficiently small we provide a consistency indicator of journals and it is a single number instead of all orders of moments.

When we ask the question of reliability of applying JIF or certain journal indicator θ to individual papers, we wish the probability that a paper with larger value of θ receives more citations than a paper with smaller θ is sufficiently high,

\[ Pr \left( c_{p(\theta_1)} > c_{p(\theta_2)} | \theta_1 > \theta_2 \right) \sim 1 \]  

(1)

Here p(θ_1) refers to one paper with indicator value θ_1 and is the received citations of that paper. When this θ refers to JIF, clearly it is not possible to guarantee the above condition. In fact, if we assume the number of received citations follows Gaussian distribution, denoted by N (x; μ, δ) then this probability can be calculated based on JIF and standard deviation via

\[ Pr \left( c_{p(\theta_1)} > c_{p(\theta_2)} | \theta_1 > \theta_2 \right) = \int_{-\infty}^{\infty} dx_2 \int_{x_2}^{\infty} dx_1 N_1 (x_1; \mu_1, \delta_1) N_2 (x_2; \mu_2, \delta_2) \]

, thus one needs not only the mean μ_2 but also the variation δ_2. Thus in this case, θ needs to include both μ and δ. When the normal distribution assumption is not valid, we then need even more information, which in principle requires all orders of moments but not only the first two, to determine the distribution function.

In this work, instead of using all orders of moments as indicators, we want to still use only JIF. For a journal, we ask the question of how large K is large enough to make Δ(K) sufficiently small so that JIF (μ(1), thus also μ(K) since μ(K) = μ(1)) become a good indicator of these K-sample from the journal. This is to ask what are the values of sufficiently large K that the following expression is more or less true,

\[ Pr \left( c_{p(\theta_1)} (K_1) > c_{p(\theta_2)} (K_2) | \theta_1 > \theta_2 \right) \sim 1 \]

(2)

which means that instead of individual papers, whether or not a set of K_1 papers with indicator value θ_1 from journal 1 can often have large received citations than another set of K_2 papers.
with indicator value $\theta_2$ from journal 2. $c_{p(\theta)}(K)$ means the received citation of a set of K papers with indicator $\theta$. This relies on the fact that $\Delta(K)$, variation of the K-sample, decreases with K so that the overlapping between $c_{p(\theta_1)}(K_1)$ and $c_{p(\theta_2)}(K_2)$ becomes very small.

In the method section, we will verify that it is a fact and also propose an explicit definition of criteria of `$\Delta(K)$ is sufficiently small'. Then, we report values of such K for a set of journals in the results section. We find that for some journals this K is very large. In this case, our calculation suggests that JIF is not representative for individual papers of the journal but it is valid to evaluate a set of K papers from that journal as a whole. Sometimes this K is even comparable to the size of the journal P, $K \sim P$. In this case, our calculation suggests that JIF probably not representative for the whole journal. For others, we find this K is much smaller than N, $1 \sim K \ll P$. In this case, our calculation suggests that JIF is reasonable to be regarded as indicators of a set of K papers from the journal. In applying this K to evaluating journals, first we need to perform the above comparison between 1, K and P to know better about the journal and then if one really insist on ranking journals, we would like to suggest the following: ignore those journals that $K \sim P$; and then for the other journals, order them according to JIF; at last for those with close values of JIF, the ones with smaller K should be favored, since keeping a very consistent impact of papers published in a journal by itself is a merit.

Our current work can easily be extended to studies on other journal-level metrics. In fact, whenever we worry about the reliability of representing a set of numbers using its average, no matter whether the underlying distribution is Gaussian or not, we can always try to calculate this large enough size K of the set. When the set has very consistent data then such K will be small and when it is highly fluctuating then K will be large. Furthermore, when K is close to the size of the set then it will be questionable to represent the whole set with its average. Therefore, the value of this K can be used to estimate whether or not the size of the data is large enough if one wants to evaluate the set by its average.

**Definition of $\kappa$**

Given a series of journals $\{j = 1, 2, \ldots, N\}$ and the citations for individual papers $\{c_{i,j}\}$, $p = 1, 2, \ldots, P_j$ in these journals during a fixed citing time window, the average citation per paper for an individual journal can be calculated as follows:

$$C^j = \langle c_{i,j}^j \rangle_p \quad (3)$$

The variance of $C_j$ of this set of N journals is as follows:

$$\delta_{1, \ldots, N} = \left\langle \left\langle C^j \right\rangle_j^2 \right\rangle_j - \left\langle \left\langle C^j \right\rangle_j \right\rangle_j^2 \quad (4)$$

This equation represents the variance in impact among the chosen N journals. If we do a bootstrap sampling with size K with replacement (every sample is put back to the set immediately after being sampled) and generate such K-sample L times for each journal, then we can get the variance of the average citation of the K samples as follows:

$$\Delta^j(K) = \left\langle \left( \frac{x^j_1 + x^j_2 + \cdots + x^j_K}{K} \right)_L^2 \right\rangle_L - \left( \left\langle \left( \frac{x^j_1 + x^j_2 + \cdots + x^j_K}{K} \right)_L \right\rangle_L \right)^2 \quad (5)$$
which represents the variance of impact within the \( K \) samples extracted from journal \( j \). We note that this sampling process and calculation of variance is similar to the concept of stability interval of the Laiden Ranking (Waltman et al., 2012).

\( \kappa_j \) is defined as the minimum \( K \) subjects to \( \Delta^j(K) \leq \delta \), that is demonstrated as follows:

\[
\kappa_j = \arg\min_K \Delta^j(K) \leq \delta_{1,\ldots,N}. \tag{6}
\]

The right hand side is just one choice of sufficiently small variation. Here we take it to be the variation among the chosen \( N \) journals. The depends on the choice of these \( N \) journals. It can be defined differently. Considering the meaning of \( \Delta(K) \) and \( \delta(1,...,N) \), the definition of \( \kappa \) is straightforward: the smallest size of papers ensuring that the variance of average impact within a journal is less than or equal to the variance between the chosen \( N \) journals. The above definition can also be explained in Fig. 1. Consider the case that two journals are chosen, \( N = 2 \) and two journals having largely overlapped distributions of number of received citations of their papers. Even when \( \mu_1 > \mu_2 \), it is hard to say that papers from one journal are of better impact than the other since \( \Delta^1(1) \sim \delta_{(12)} \), thus \( \Delta^2(1) \sim \delta_{(12)} \), thus \( \Delta^1, \Delta^2 \ll \delta_{(12)} \).

In this way, the overlapping between random variables \( c_{p(\theta_1)}(K_1) \) and \( c_{p(\theta_2)}(K_2) \) is very small so that Eq. 2 holds.

Next we confirm from empirical tests, although it can also be shown theoretically, that the bootstrap technique in statistics guarantees that for large enough \( L \), the \( K \)-sample variance \( \Delta^j(K) \) decreases with larger \( K \) for a distribution function with finite variance or for any set of empirical data with large variance that \( \Delta^j(K) \) remains finite (Wasserman, 2004). This can be verified in Fig. 2. Using real data from a set of journals, we verify that \( \Delta^j(K) \) indeed decreases with \( K \) as shown in Fig. 2. Since as shown in Fig. 2 \( \Delta(K) \) is very close to \( K/\Delta \), our definition of \( \kappa_j \) can be regarded as a conversion to \( \Delta^j \) as in \( \kappa_j^j = \frac{\Delta^j}{\delta} \).

![Figure 1: Definition of \( \kappa \) for two journals initially have largely overlapped distribution.](image)

(a) Originally two journals have largely overlapped distributions such that \( \Delta^1, \Delta^2 \ll \delta_{(12)} \).

(b) After taking \( \kappa_1 \) (\( \kappa_2 \)) samples from journal 1 (2), the resulting distribution functions
are well separated such that $\Delta^1 (\kappa_1), \Delta^2 (\kappa_2) \ll \delta_{(12)}$. (c) We can take $\kappa_{1,2} = 1$ for two journals whose distributions are already well separated such that $\Delta^1 (1), \Delta^2 (1) \ll \delta_{(12)}$.

The fact that our definition of $\kappa_j$ can be regarded as a conversion directly from $\Delta^j$, the standard variation of journal $j$, without involving bootstrapping at all makes tectonically the bootstrap procedure irrelevant to our discussion. However, logically the bootstrap procedure is still necessary since it tells us that we may now regard the $K$-sample average as a mean to evaluate the $K$ papers from the journal $j$ as a whole and consider average and variation of various choice of $K$ papers. Furthermore, in pure theoretical consideration, this bootstrap procedure is also necessary since, in theory, we might need to discuss average and variance of various $K$ samples from a heavy-tail distribution function, such as a power-law function, of citation counts, in which case, $\Delta (K) = \frac{\Delta}{K}$ might not hold anymore.

From the definition of $\kappa$, we can see that $\kappa$ is both determined by the selected journals and the citation distribution of each journal. A significant variance in the numbers of received citations of the selected journals results in a large $\Delta$, thus $\kappa$ will be small and the journals with papers yielding higher consistency in the numbers of received citations will have smaller $\kappa$.

Figure 2: Empirical results verify that $\Delta(K)$ decreases with $K$. Here the horizontal line corresponds to the value of group $\delta$. We can also see that $\Delta(K)$ is very close to $\Delta/K$. Here we use $L = 4000$.

3. Results

3.1. Data

We extracted papers published in 2010 and their corresponding citation count between 2010 and 2012 of 25 journals from Web of Science. The journals selected are both multidisciplinary and subject-specific in scope, and cover a wide range of average citations per paper. Names and other information of selected journals can be found in Table 1. We chose these 25 journals to simply implement and test our ideas. Therefore, we have selected multidisciplinary and specialized journals related to medical studies. Only publications of the Web of Science document type “article” were included in the data collection.

3.2. Pairwise comparison
JIF (or journal average citation) is often used to assess the performance of individual researchers. A researcher with a single paper published in a higher JIF journal is customarily thought to possess greater academic influence than a researcher with a paper published in lower JIF journals. This claim is established in low confidence, even if it may be true. For example, as shown in Fig. 3(a), the citation distribution of Nature and Nat Cell Biol shows a high degree of overlap, rendering the citation distribution indistinguishable between individual papers although the average number of citation in Nature is higher.

Taking these two journals as our test set of journals, and using citations of papers in the two journals, we calculate the minimum number of representative papers $\kappa$ for each of the two journals, as according to Eq. 5 and Eq. 6, where $\delta$ is determined as the variance of the average citation amount $C$ of the respective journals. We found that $\kappa$ for Nature and Nat Chem Biol is $\kappa_{NCB}^N = 38$ and $\kappa_{NCB}^N = 9$ respectively. Here, we introduce a notation $\kappa_c^r$ to denote the minimum number of representative papers of journal $r$ when compared against journal $c$. The large value of $\kappa_{NCB}^N = 38$ associated with Nature is partially due to the highly heterogeneous citation distribution of Nature, in addition to the comparability between the average citation values of the two journals.

The distributions of average citations of $\kappa$-sample, which is the calculated average of a random selection of $\kappa$ papers from a journal are shown in Fig. 3(b). A clear distinction between the average citations of Nature and Nat Cell Biol emerges, and the intra-journal variance shrinks sharply; this is indicative of high confidence in correctly distinguishing the journals based only on citations.

Complete results of the pairwise comparison are shown in Fig. 4. Each small square at position $(r, c)$ represents the value of $\kappa_c^r$ in logarithmic scale for the row journals when compared against the column indexed journals; the color red indicates that the average citation $C_r$ of the row journal is larger than that of the column journal, otherwise, it is blue. We observe that some journals, e.g., Nat Struct Mol Biol ($r = 10$) and PNAS ($r = 12$), yield consistently low $\kappa$ when compared to other journals with $C$ ranging from high to low; this indicates that the two journals exhibit a high degree of consistency in their publications. Conversely, JAMA ($r = 8$), as compared with the remaining journals, has a relatively high $\kappa$. One possible reason for this is the highly skewed citations of the papers published in 2010. For most journals, the darkest squares are along the quasi-diagonal ($\pm 1$ shift), which is indicative of the journal pairs along the quasi-diagonal having similar average citations, thereby resulting in a small $\delta$, and thus a large $\kappa$. This is understandable since the two neighboring journals are not significantly distinguishable according to their JIFs, and if we want to distinguish them, a substantially large $\kappa$ is required, in some cases even much larger than their publication size. Furthermore, the white-dominated portion in the lower left of Fig. 4 represents a large difference in the average citation $C$ between the compared pairs of journals. The selected journals belong to various disciplines that may have significantly different citation patterns. However, for simplicity, we did not include any disciplinary normalization method, although it should be noted that whenever possible a disciplinary normalization should always be used (Radicchi & Castellano, 2008, Waltman, 2016) and we will do so in future works which will be more results oriented than the current method oriented work. Thus, in principle, here interdisciplinary journal Nature and Science should not be compared with Nat Cell Biol.

From the above pairwise comparison, each journal $i$ gets one value of $\kappa_c^j$ when compared with a journal $j$. Thus, in total, each journal gets $N - 1$ such numbers, thereby disallowing the
assignment of only one number to a journal. In order to define one number as the minimal amount of representative papers of each journal based on the pairwise comparison, we then divide the entire set of papers corresponding to all journals into two sets, one set being papers from journal j and one set the remaining papers.

Figure 3: Box-plot for the citation distribution of *Nature* and *Nat Cell Biol*. (a) The original citation distribution for Nature (left purple) and Nat Cell Biol (right green) are shown in logarithm because of their highly skewed distributions. The vertical-axis ‘citation + 1’ is used in consideration of papers without citations to avoid log (0). The five horizontal lines in each box represent the maximum, third quartile, median, first quartile and minimum amount of citations, respectively. The black plus markers represent the average citation. (b) Distribution of the bootstrap K-sample average citation when the typical minimum number $\kappa = 38$ for *Nature* and $\kappa = 9$ for *Nat Cell Biol*, respectively.

Next, we compare the two sets and define a single indicator of the minimal representative size of journal j under pairwise comparison to be $\kappa^j_R = \kappa^{\text{rest}}_j$. Values of $\kappa^j$ for all the journals studied in this work are provided in Table 1. It was worth noting that $\kappa^j$ is similar to its publication size for *NAT CHEM BIOL* and is even much larger than the publication size for *JAMA*. 
Figure 4: Heatmap of the pairwise compared $\kappa$. Journals are sorted in descending order based on their average citation (same order as Table 1). The darkness of each small square represents the value of $\kappa$ in logarithmic scale for its corresponding row journal $r$ when compared with its column journal $c$: red squares indicate the average citation of the corresponding row journal is larger than that of the column journal, and vice versa for blue squares. For example, the darkness of the square (3, 7) represents the value of $\kappa^3_7 = 38$ of Nature ($r = 3$) when comparing with Nat Cell Biol ($c = 7$).

3.3. Group Comparison

Next, let us extend this comparison between two journals or two sets to a comparison of many journals. Consider all of the journals from which we collected individual paper citations as a group $J$. This is applied in Eq. 5 and Eq. 6 to calculate $\kappa$ for each journals in the set, where $\delta_j$ is calculated as the variance of $C$ for all the journals in group $J$. This $\kappa^G_j$ indicates the minimum number of papers required to represent the impact of journal $j$ while also yielding distinguishability among the group of journals $J$.

The original citation distributions of all journals in our set are shown in Fig. 5(a). Although the journals are ordered based on their average citation $C$, there are high degrees of overlaps between the left-most journals and the right-most journals. Such overlaps prevent us from using the JIF to evaluate individual researchers and individual papers. However, after calculating $\kappa^G_j$, the $\kappa^G_j$-sample average become well separated as can be seen from Fig. 5(b). Thus, this may facilitate distinguishing of a single journal from the entire set of journals.

The list of values of $\kappa^G$ for each journal is shown in Table 1. For Nature, 11 papers are needed to represent its impact as compared with the remaining journals, and for Science, 6 papers are enough. Note that when Nature and Science are compared directly, the pairwise $\kappa$ for each are $\kappa^S_N = 173$ and $\kappa^S_N = 90$, respectively, which are different from their respective group $\kappa^G$ values. For journals with $\kappa = 1$, e.g., Mol Psychiatr and Alcohol, their sizes are small and they exhibit consistency in the impact of publications when compared with the remaining journals in the entire set; however, the consistency does not hold true when performing direct comparisons between individual journals.

The details of all chosen journals are shown in Table 1. We note that in most cases, the value of $\kappa^p_j$ of journal $j$ is larger than $\kappa^G_j$. This is particularly seen for example of Nat Chem Biol and JAMA. Yet, both $\kappa^p$ and $\kappa^G$ are often much smaller (except Nat Chem Biol and JAMA) than the corresponding publication size. This implies that JIF for those journals is representative in the sense that if approximately $\kappa^p_j$ or $\kappa^G_j$ papers were randomly chosen from the journal $j$, the average is close enough to JIF. However, when $\kappa^{p,G}_{j} \gg 1$ it also implies that it is not a good idea to use JIF to evaluate individual papers of those journals. Or when $\kappa^{p,G}_{j} \sim P_j$ close to the size of the journal, then it is not a good idea to use JIF even to evaluate this journal, for example, in the case of JAMA. Therefore, we think that calculating this $\kappa$ and comparing it against 1 and $P_j$ tells us valuable information on the reliability of using JIF to evaluate journal J.
Figure 5: Box-plot for the citation distributions of all journals in Table 1. (a) The original citation distribution for all journals considered in this work. The journals appear in descending order according to their average citation $C$. (b) The distribution of $\kappa^G$-sample average citations, where $\kappa^G$ of each journal is calculated according to Eq. 6 and listed in Table 1. Compared to (a), the distributions in (b) are in much narrow ranges. The five horizontal lines in each box represent the maximum, third quartile, median, first quartile and minimum amount of citations, respectively.

4. Conclusion and Discussion

Quantifying the performance of scientists and impact of papers is a trending and controversial topic. Our above analyses demonstrate that, although it is not appropriate to utilize a journal-level metric, such as JIF, to evaluate individual researchers and papers, it is still a reasonable metric and effective in practice to evaluate a group of papers if the sizes of such groups are significantly large. In this work, we proposed the following three definitions: pairwise $\kappa_{ij}$, one-against-the-rest $\kappa^P_j$ and group-based $\kappa^G_j$. We find that $\kappa^P_j$ and $\kappa^G_j$ are often much smaller than the publication size of the journal $j$. Knowing $\kappa_j$, the minimum number of representative papers of journal $j$, means that if JIF is part of the indicators according to which journals are going to be evaluated, then $\kappa_j$ must be much less than the size of the journal, $\kappa_j << P_j$.

For journals, often this condition can be easily satisfied since it is not that hard to increase the number of papers of a journal. However, it might become a problem when using JIF-like indicators to rank universities or departments since for some of them we may find that their number of total publications might not be much larger than their necessary size of representative samples. In practice, often there is a hand-picked threshold, like universities with less than 500 publications in the dataset are discarded (Bornmann et al., 2014). Having a definition of the necessary size of representative samples in this case helps to draw a line, although it might be a very fuzzy line, between the large enough universities/departments and the not so large ones according to their own consistency levels. For example, some small universities, which would be discarded previously, can still be taken into consideration if they have high consistency levels.

It should be noted that, according to the definition, journals with papers yielding greater consistency in the number of received citations have a smaller $\kappa$. Therefore, $\kappa^P$ and $\kappa^G$ can also be regarded as journal-level indicators of consistency. When ranking journals both JIF and these $\kappa$ should be considered.
However, we admit that like percentile ranks the value of \( \kappa \) depends on the choice of the entire set of journals and thus their values are more complicated than the JIF. In future studies, one might want to divide the journals into different groups according to for example JIF and then do the group comparison \( \kappa_G \) within the predefined groups since it is not very possible that a journal with very low JIF will like to be compared against a journal with very high JIF. In the current work, we treated all journals as a whole group.

As shown in Table 1, many journals possess a minimum number of representative papers equal to one. This means that those journals have small within-journal citation variations while the inter-journal group variation is much bigger. We suppose one reason for this is related to the current set of chosen journals, where the average citations of the journals are relatively already well separated. This results in the value of the group variation \( \delta_J \) being large while some journals have small single-paper variations \( \Delta(K = 1) \) just as illustrated in Fig. 1(c). In practice, this problem may be overcome by choosing a sufficiently large group of journals to cover all the journals of interest to the researchers, or by dividing the data set into subsets with a smaller group variation \( \delta' \) within the subsets. We are planning to collect a larger set of journals and perform a more detailed study.

Lastly, we would like to emphasize again that even if we do find \( \kappa_j \ll P_j \) and \( \kappa_j \approx 1 \) for a journal \( j \) when it comes to evaluating individual papers from the journal, quantitative indicators should only be supplementary to peer review, just that we might feel more comfortable in using these indicators.

Besides the investigation on minimum representative size of journals, we think our explicit mathematical statement of ‘reliably evaluating single papers according to indicator \( \theta \)’, i.e. Eq. 1, \( P_r \left( c_{p(\theta_1)} > c_{p(\theta_2)} \mid \theta_1 > \theta_2 \right) \sim 1 \), is also meaningful and it may stimulate further investigations on applicability of all journal-level indicators.

Table 1: \( \kappa \) values for the medical journals considered here. For each journal, its name, number of papers (\( P \)), average citations per paper (\( C \)), and minimum number of representative papers (\( \kappa_C \) and \( \kappa_P \) ) are shown.

<table>
<thead>
<tr>
<th>ID</th>
<th>Journals</th>
<th>( P )</th>
<th>( C )</th>
<th>( \kappa_C )</th>
<th>( \kappa_P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LANCET</td>
<td>242</td>
<td>72.52</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>NEW ENGL J MED</td>
<td>310</td>
<td>69.63</td>
<td>17</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>NATURE</td>
<td>877</td>
<td>69.05</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>CELL</td>
<td>288</td>
<td>57.32</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>SCIENCE</td>
<td>799</td>
<td>56.98</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>NAT MED</td>
<td>135</td>
<td>43.79</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>7</td>
<td>NAT CELL BIOL</td>
<td>128</td>
<td>43.28</td>
<td>3</td>
<td>48</td>
</tr>
<tr>
<td>8</td>
<td>JAMA-J AM MED ASSOC</td>
<td>218</td>
<td>38.33</td>
<td>21</td>
<td>1236</td>
</tr>
<tr>
<td>9</td>
<td>NAT CHEM BIOL</td>
<td>96</td>
<td>28.31</td>
<td>1</td>
<td>102</td>
</tr>
<tr>
<td>10</td>
<td>NAT STRUC MOL BIOL</td>
<td>209</td>
<td>25.11</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>11</td>
<td>MOL PSYCHIATR</td>
<td>91</td>
<td>22</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>12</td>
<td>P NATL ACAD SCI</td>
<td>3759</td>
<td>19.98</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>CURR BIOL</td>
<td>341</td>
<td>17.07</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>BMJ-BRIT MED J</td>
<td>245</td>
<td>14.33</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>
Acknowledgments
We thank Mr. Jianlin Zhou, Per Aglgren, Ludo Waltman and Lawrence Smolinsky for valuable discussion. This work was supported by NSFC under Grant No. 61374175 and the Fundamental Research Funds for the Central Universities.

References


Do Mathematicians, Economists and Biomedical Scientists Trace Large Topics More Strongly Than Physicists?

Menghui Li¹ Liying Yang² Huina Zhang¹ Zhesi Shen³ Chensheng Wu¹ Jinshan Wu³

¹limhxd@126.com
²National Science Library, Chinese Academy of Sciences, Beijing 100190, P. R. China
³School of Systems Science, Beijing Normal University, Beijing 100875, P.R. China

Abstract
In this work, we extend our previous work on largeness tracing among physicists to other fields, namely mathematics, economics and biomedical science. Overall, the results confirm our previous discovery, indicating that scientists in all these fields trace large topics. Surprisingly, however, it seems that researchers in mathematics tend to be more likely to trace large topics than those in the other fields. We also find that on average, papers in top journals are less largeness-driven. We compare researchers from the USA, Germany, Japan and China and find that Chinese researchers exhibit consistently larger exponents, indicating that in all these fields, Chinese researchers trace large topics more strongly than others. Further correlation analyses between the degree of largeness tracing and the numbers of authors, affiliations and references per paper reveal positive correlations – papers with more authors, affiliations or references are likely to be more largeness-driven, with several interesting and noteworthy exceptions: in economics, papers with more references are not necessarily more largeness-driven, and the same is true for papers with more authors in biomedical science. We believe that these empirical discoveries may be valuable to science policy-makers.

Conference Topic
Scientometrics

Introduction
How researchers choose their research topics has received sustained attention (Busch et al., 1983, Diamond, 1994, Gieryn, 1978, Foster et al., 2015). This will not only directly affect scientists’ output and recognition, but also indirectly affect the science itself. This problem has been widely investigated from sociology of science (Merton, 1957, 1968, Latour, 1987), philosophy of science (Kitcher, 1995, Strevens, 2003, Kleinberg & Oren, 2011, Bikard et al., 2015, Boyer-Kassem & Imbert, 2015), and new economics of science (Dasgupta & David, 1994, Stephan, 1996). Quite often previous investigations on this issue are rather qualitative. In this work, we want to study this empirically based on data.

Researchers may determine their research topics primarily according to their research interests, their perceived potential in making progress on the topics or simply by the largeness of the topics, or some combination of all these factors (Foster et al., 2015). Here, largeness refers to how many publications are on the topic during a time interval denoted as \([t_0, t]\). For example, based on knowledge network of chemical reactants, Rzhetsky et al. explore the probability of selecting a pair of molecules as a function of the importance (represented by degree) and the difficulty associated with combining them (measured by network distance). It is found that biomedical scientists prefer to explore the local neighborhood of central, important (i.e., higher-degree) molecules in biomedical chemistry (Rzhetsky et al., 2015). In this work, we focus on the effect of largeness of topics on scientists’ choices of topics. It has been found that the scientists tend to trace large topics in physics (Wei et al., 2013), and in environmental science (Grandjean et al., 2011).

“The rich get richer” or Matthew effect is a common social phenomenon (Barabasi & Albert, 1999, Price, 1976, Simon, 1955). Matthew effects in science have been investigated
also in scientometrics, for instance, on scientists’ credit (Merton, 1968) and on citations (Bonitz et al., 1997, Biglu, 2008). Only a few previous studies are on scientists’ choice of topics in a certain discipline (Wei et al., 2013, Rzhetsky et al., 2015, Grandjean et al., 2011). Thus, in this paper we ask: Does tracing large topics differ across disciplines? What is the difference of degree of largeness tracing among different countries? To what degree is the intensity of largeness tracing relevant to properties such as the number of authors, references and affiliations of articles? We hope that discoveries from this study will provide valuable information for scientific policy makers especially on the issue of funding and evaluation.

The question of hot topics, prominent topics, or research fronts itself has also been widely investigated in scientometrics (Boyack & Klavans, 2010, Small et al., 2014, Upham & Small, 2010, Chen, 2004, Cozzens et al., 2010, Klavans & Boyack, 2016). Research fronts are generally defined as the areas attracting the most scientific interest in a period of time, especially before publication size of the considered field becomes really big and the field is clearly under exponential growth. Therefore, the studies of hot topics conventionally refer more to identifying emerging hot fields. We will refer these studies as studies on emerging hotness. One might use the number of received citations of papers on a topic (Boyack & Klavans, 2010, Small et al., 2014, Upham & Small, 2010, Chen, 2004, Klavans & Boyack, 2016), or simply use the number of publications (or scientists) on a topic (Cozzens et al., 2010) for this purpose. In this work, we do not need to identify emerging research fronts and our use of hotness or largeness is more like the current size (in terms of number of publications) of the fields. We call the hotness in this sense the realized hotness, or simply largeness.

Furthermore, for each of the four disciplines that we study, which are mathematics, physics, economics and biomedicine, there is an established hierarchical classification system of topics and all the publications under our investigation have been annotated with such a hierarchical code from the system. Therefore, we do not even need to do classification or clustering of publications to identify topics. In this work, articles with same code at a certain level are considered to be on the same topic. Both the number of papers and the number of received citations of the papers on a topic can be used to quantify the largeness of the topic. However, these two quantities are not really independent: the number of citations more or less follows a power law relation with respect to the number of papers (Katz, 1999). In the following analysis, we use only the number of publications belonging to a topic as a measure of the largeness of the topic as for example in Rzhetsky et al. (2015), Grandjean et al. (2011). In principle, one can also consider how new publications are attracted by received citations of topics, but this will be the topic of another investigation. Thus, using established hierarchical classification systems of topics and classifications of papers in these systems, our definition of large topics in this paper is much simpler than that of research fronts: We only need to count how many papers belong to a topic at a given level. The more papers, the larger the topic and the bigger the realized hotness. Using the established hierarchical classification systems of topics is not ideal since new topics emerge constantly. One may use various clustering methods based on citation relation among publications to establish classification systems and identify topics (Subelj et al., 2016, Boyack & Klavans, 2010, Klavans & Boyack, 2016). However, in this work, we will use the established coding systems based on controlled vocabularies for the above four disciplines.

Using the above measure of largeness, in this work, we investigate the degree of largeness tracing in mathematics, economics and biomedical science, and then perform a comparison among them and also between these three fields and physics. Intuitively, it might be expected that mathematicians would be more likely to choose their topics of investigation according to their scientific interests and the scientific value of the questions, partly due to the fact that mathematicians intend to work individually and partly since often mathematicians claim so.
One might also guess that it is possible that biomedical scientists choose research topics mainly according to medical values of problems rather than the largeness of topics. Economists might choose their topics according to their urgency to the current economic situation rather than their largeness. Here, we empirically examine whether or not this is the case.

Aside from pure curiosity, such a comparison among various fields might have value in the study of science policy. For example, we wish to examine the correlation between the impact of papers and their degree of largeness tracing. As researchers, we would prefer to see that less largeness driven papers have a greater impact. For science policy makers and administrators of universities, funding agencies and institutes, such a correlation, whether positive or negative, could potentially provide guidance regarding their duties, as one of their goals is to seek strategies to improve the scientific impact of the academic units under their administration. We would also like to examine whether larger teams tend to produce more value-driven papers or more largeness-driven papers. An answer to this question might have strategic policy value regarding how large research teams should be supported.

In our previous work, using data obtained from the American Physical Society (APS) concerning publications in APS journals, we investigated and confirmed the occurrence of largeness tracing in physics (Wei et al., 2013). In this work, we extended this analysis to the fields of mathematics, economics and biomedical science based on publications from the datasets of the Mathematical Review (MR), the Journal of Economic Literature (JEL) and PubMed, respectively. Each of these databases uses a field-specific classification scheme – the Physics and Astronomy Classification Scheme (PACS) for physics, the Mathematics Subject Classification (MSC) for mathematics, the JEL Classification Codes (JEL) for economics and the Medical Subject Headings (MeSH) for biomedical science – and classifies each paper into one or more categories, which we call fields or subfields when the categories become finer and finer. When more specific information is needed, we use the following notation, \( n \)-digit topics, such as 1-digit topics, 2-digit topics etc. to denote the fields and subfields at the level of \( n \)th number/letter of the corresponding classification system. In this sense, we call the 0th level topics discipline, namely here Physics, Mathematics, Economics and Biomedical science. We test how likely a newly published paper is to be in a large field, where the largeness of each field is measured simply by the current number of published papers in that field accumulated starting from some early years about which we still have publication data. Further details on our methods and data can be found in section 2, where we also define a quantity called the relative contribution ratio, \( R_c(k) \), to describe the extent to which researchers from a given country contribute to fields of size \( k \) compared with the overall contributions of this country to academia. The exponent \( \alpha \) and the relative contribution ratio \( R_c(k) \) are the statistics that we analyze throughout the remainder of the paper.

**Methods and Data**

Given a set of newly published papers in a time window at time \( t \) with size \( \Delta t \) ([\( t, t + \Delta t \)]), we first record in which fields each paper of them belongs to. We also record the current number of published papers in each field till time \( t \) (papers published between \( t \) and \( t + \Delta t \) are not included) starting from some early years \( t_0 \) about which we still have publication data. Then by combining the above two pieces of information, we count how many of the newly published papers appear in fields of size \( k \), which we denote by \( p_k(t) \). We are interested in knowing how \( p(k) \) depends on \( k \). According to various previous studies (Barabasi & Albert, 1999, Price, 1976, Simon, 1955, Wei et al., 2013, Rzhetsky et al., 2015, Grandjean et al., 2011), often there is a power law \( p(k) \sim k^\alpha \). In principle, one can extract this component \( \alpha \) by directly fitting \( p(k) \) with respect to \( k \).
However, there might be more than one topic with size \(k\) (we denote number of size-\(k\) topics as \(n(k)\)) and often \(n(k)\) also follows a power law with respect to \(k\), \(n(k) \sim k^{-\gamma}\). Considering that for each single size-\(k\) topic \(p(k) \sim k^\alpha\) (this exponent \(\alpha\) is what we really want) and there are \(n(k)\) such topics, we will have \(p(k) \sim k^\alpha n(k) \sim k^{\alpha-\gamma}\). Therefore, if we directly fitting \(p(k)\) to \(k\), we would have to know the \(\gamma\) of \(n(k)\). In order to get rid of this additional parameter, we define

\[
T(k) = \frac{\sum p(t)}{\sum \tau(t)}
\]

The normalization is to make \(T(k)\) more like a distribution function. It has been found that in physics this curve is close to a power law (Wei et al., 2013). We then fit this curve of \(T(k)\) to a power law

\[
T(k) \sim k^\alpha
\]

and refer to the exponent \(\alpha\) as the degree of largeness tracing. In fact, to smooth the curves, we use the integral of \(T(k)\) over \(k\), which we call the accumulated distribution function

\[
K(k) = \int_0^k T(x)dx
\]

Consider a special case where the likelihood of each current paper leading to a new publication is the same, then \(p(k) \sim n(k) k\) and \(T_k \sim k\). This is a constant ‘birth rate’ model, which is unlikely to be true since some papers might lead to a lot of new papers while many other papers will not inspire any further publication. Thus \(\alpha = 1\) is an interesting special case. Another special case is \(\alpha = 0\) which means that large fields and small fields are equally attractive or inspiring for researchers when they decide to work on which topics. What are the empirical values of the exponent \(\alpha\) for various disciplines, or for authors from various countries, might reveal valuable and interesting information about the disciplines and the countries. Therefore, in the following, besides comparing empirical values of \(\alpha\) against 1 and 0, we will also focus on cross-discipline/cross countries comparisons, and a few other comparative studies to reveal how other corresponding factors influence largeness tracing.

In addition to the degree of largeness tracing \(\alpha\), we also define the relative contribution ratio \(R_c(k)\) for each individual country \(c\). Given a set of papers and the fields to which they belong, we calculate a quantity \(m^c_k\), which is equal to the number of papers contributed by country \(c\) in fields of size \(k\) considering all publications from \(t_0\) to \(t\) in our datasets. In counting \(m^c_k\), we use fractional counting towards countries, in the sense that a country that occurs \(y\) times among the \(x\) affiliations of a paper is counted as a contribution of \(y/x\) from that country. Using this \(m^c_k\), we define

\[
R^c(k) = \frac{\sum m^c_k}{\sum \sum m^c_k}
\]

Intuitively, the numerator represents the fraction of the contributions made by country \(c\) out of all papers in fields of size \(k\), whereas the denominator represents the fractional contribution of country \(c\) to all fields. More detailed information can be found in (Wei et al., 2013).

In our datasets, each paper is recorded as a data entry that includes the title, year of publication, subject classification code(s), author(s), affiliation(s), reference(s) and number of received citations. The classification schemes in these datasets are all hierarchical. In this work, we use the finest level of each classification system. Articles with the same code are classified into the same topic. Moreover, we regard each paper as one unit of contribution to each assigned topic even when it is assigned to more than one topics. It is not exactly accurate but reasonable for now and we should consider how to evaluate the contribution of each paper.
to their topics more accurately in future studies. It is found that the size of topics follows a skewed distribution (Wei et al., 2013).

The physics dataset is a collection of all papers published in APS journals from 1976 to 2013. Here, we consider only those research papers, e.g., articles, brief reports and rapid communications, that have been annotated with PACS numbers. In physics, we consider 6-digit topics, e.g., 03.65.-w (Quantum mechanics). We consider a total of 389912 papers, 5823 PACS numbers, and 1131566 classification labels.

The mathematics dataset is a collection of all papers collected by Mathematical Reviews from 1969 to 2015. Here, we consider only those research papers recorded by both Mathematical Reviews and Web of Science (WoS). The MSC codes for each paper are taken from Mathematical Reviews, and other information is from WoS. There are significant differences between different versions of the MSC codes. Thus, we study the mathematics dataset separately for three different time periods, from 1991 to 1999, from 2000 to 2009 and from 2010 to 2015. In mathematics, we consider 5-digit topics, e.g., 40C05 (Matrix methods). We have respectively 92992, 289554 and 261504 papers, 5143, 5124 and 5487 MSC codes, with 456994, 456994, 755121, 711349 classification labels for the period of 1991-1999, 2000-2009 and 2010-2015.

The economics dataset is a collection of all economics papers collected in the American Economic Association JEL database from 1970 to 2013. Here, we consider only those papers recorded by both JEL and WoS. The JEL codes are taken from the JEL database, whereas other information is from WoS. In economics, we consider 3-digit topics, e.g., D12 (Consumer Economics: Empirical Analysis). We consider 241751 papers, 1125 JEL codes and 582150 classification labels.

The biomedical science dataset is from PubMed, which covers a wide range of important journals in the biomedical and life sciences dating back to 1950. We downloaded the 2015 baseline version provided by PubMed. We consider only those papers which have been assigned with MeSH codes. MeSH codes is a classification tree and different branches might have different depths. Some branches stop at 12-digit codes such as A01.456.505.580 while others have a depth of 9-digit such as A17.360.296 or depth of 24-digit such as D27.505.696.875.360.276.210.277. In this work, we only report results on the 12-digit topics.

Figure 1: Empirical preferential attachment to large fields for newly published papers. The cumulative probability functions $K(k)$ for the (a) physics, (b) mathematics, (c) economics and (d) biomedical science datasets show various values of $\alpha$. The biomedical science dataset is from PubMed, which covers a wide range of important journals in the biomedical and life sciences dating back to 1950. We downloaded the 2015 baseline version provided by PubMed. We consider only those papers which have been assigned with MeSH codes. MeSH codes is a classification tree and different branches might have different depths. Some branches stop at 12-digit codes such as A01.456.505.580 while others have a depth of 9-digit such as A17.360.296 or depth of 24-digit such as D27.505.696.875.360.276.210.277. In this work, we only report results on the 12-digit topics.
At last, we have 21850751 papers, 6855 MeSH codes, and 123446643 classification labels. We have not yet integrated this data with WoS; consequently, many entries have no affiliations and/or no references.

Results

We first consider sets consisting of all papers published in each year in each of the three disciplines, determine the overall degree of largeness tracing in each discipline in each year, and compare the results with the degree of largeness tracing in physics. We then divide the set of papers in each discipline according to various characteristics of each paper, including whether it was published in a top journal, its country of origin, the number of authors, the number of affiliations and the number of references, and study the correlations between these features and the degree of largeness tracing.

![Figure 2: Evolution of exponents. The exponents for different years obtained by least-squares fitting of the curves from k = 10 to k = 300 for physics, mathematics and economics and from k = 10 to k = 3000 for biomedical science. These fitting ranges were so chosen because the curves deviate from straight lines at large k due to empirically the rarity of such extreme values. The standard deviations of the fitted value of \( \alpha \) are smaller than 0.01. The goodness of regressions \( R^2 \) are larger than 0.99. Unless otherwise stated, \( \Delta t \) is equal to one year for physics, mathematics and economics and one month for biomedical science. The results for biomedical science are the average values for the corresponding 12 months. In the year 2000 and 2010, there is a significant change to MSC so that the exponent cannot be calculated.](image)

Overall, scientists do trace large topics

As shown in Fig. 1, the accumulated distribution function \( K(k) \) follows a power law, namely, \( K(k) \sim k^\alpha + 1 \). All of the power-law exponents \( \alpha \) are positive, indicating that hotter fields (those that are larger in size) do attract more newly published papers. These results are qualitatively consistent with our observations in physics: generally, scientists are more likely to choose to publish in hotter subfields. The reasons may be as follows. Scientists publish their achievements with high probability by following large topics, so they may more easily achieve enough recognition to maintain a position. It is also possible that development of science has made a strong correlation between scientific value and largeness of research topics, such that the large topics are the one with most scientific value. In this case, then it is not surprising that overall scientists trace large topics. Unfortunately, it is hard to put this to a test since so far there is no good objective measure of the scientific value of topics. However, differences between the degrees of largeness tracing among the various disciplines are still informative. In particular, the exponent in mathematics is markedly larger than those in the other three fields, as is also shown in Fig. 2 for more than 20 years. This is surprising and counter-intuitive, as it is widely claimed, at least among mathematicians, that the scientific efforts of mathematicians are more independent and more interest-driven or value-driven rather than largeness-driven. The reasons for this discrepancy are not yet known to us. What is
also noteworthy is that the exponent in biomedical is less than those in other fields, indicating probably that biomedical scientists are more interest-driven rather than largeness-driven.

**Figure 3: Relation between largeness tracing and cited number.** The preferential attachment exponents $\alpha$ increase slightly with respect to the cited number $N_{\text{Cited}}$.

*Influence of journals on largeness tracing: Papers in top journals are less largeness-driven*

We also like to know whether, on average, largeness-tracing papers have less impact on academia as a whole. To this end, we can investigate the correlation between the exponent $\alpha$ and the number of received citations. As shown in Fig. 3, doing so reveals no clear-cut pattern, except that in mathematics, there is an overall trend of an increasing value of the exponent with an increasing number of received citations. However, it is generally accepted that papers that are published in top journals are more likely to be original and have a higher impact. Thus, it is possible that they are less largeness-driven. Here we compare the degree of largeness tracing among papers published in top journals with that among papers published in other journals.

We classify the papers into two subsets, i.e., those published in top journals, which are chosen from our databases by disciplinary experts to choose only the highest recognized ones, and those published in other journals, and for each subset, we calculate the exponent $\alpha$. As shown in Fig. 4, overall, the exponents $\alpha$ for papers in top journals are smaller than those for papers in other journals in all four disciplines. This finding indicates that publications in top journals are indeed less largeness-driven. The values of the exponent $\alpha$, as indicators of the degree of largeness tracing, are 0.59 (0.73) in physics, 0.68 (0.92) in mathematics, 0.52 (0.64) in economics and 0.33 (0.45) in biomedical science for top (other) journals. We also apply the two-sample Kolmogorov-Smirnov test (KS test) to the sets consisting of the recorded $k$ values of each paper at the time of its publication in top and other journals, and the results of the KS test indicate that for all these disciplines, it is safe to reject the hypothesis that these two sets of $k$ values are drawn from the same underlying distribution. This supports the discovered difference between the exponent $\alpha$ that it is possible that the degrees of largeness tracing of top journal papers are statistically different from that of other papers.

*Comparison of the degrees of largeness tracing among various countries: Chinese scholars are more largeness-driven*

General public and science policy-makers have a great interest in the research performance of various countries. Here, we compare the degrees of largeness tracing among various countries. We classify the papers into subsets according to the countries that appear in their affiliations. When the authors are from more than one country, the paper is included in the set corresponding to each country. The absolute contribution ratios are listed in table 1. Note that in economics, the contributions from Japan and China are very low. As we will see later, this leads to extremely large fluctuations and, consequently, less reliable observations.
As seen from Fig. 5(a-c), the exponents calculated for China are consistently larger than those for the other three countries in physics, mathematics and economics. Note that the value of Chinese scholars on mathematics is even larger than 1. Recall that $\alpha = 1$ means constant ‘birth rate’ model, where the likelihood of inspiring/attracting/leading to a new publication is the same for each current paper. With this $\alpha > 1$, it implies that for Chinese mathematicians it is even possible that larger fields are disproportionately more attractive than the other fields. We do not want to speculate further on why it is the case in China. However, since this observation that Chinese scholars have larger $\alpha$ is there in all the three disciplines (In Economics, it is still larger but not that much larger than the $\alpha$ of others countries), other researchers should look into possible systematic reasons behind it. To further characterize the differences among these countries, we consider the relative contribution ratios $R_c(k)$.

Figure 4: Comparisons of largeness tracing between top journals and other journals: (a) physics, in which the top journal is Physical Review Letters; (b) mathematics, in which the top journals are Inventiones Mathematicae, Annals of Mathematics, Acta Mathematica, and the Journal of the American Mathematical Society; (c) economics, in which the top journals are the American Economic Review, Econometrica, the Journal of Political Economy, the Quarterly Journal of Economics, and the Review of Economic Studies; and (d) biomedical science, in which the top journals are Nature, Science, the Proceedings of the National Academy of Sciences of the United States of America, Cell, the New England Journal of Medicine and Lancet.

As shown in Fig. 5(d-f), in all three of these disciplines, the relative contribution ratio $R(k)$ of China is smaller in cold fields (for small $k$ values) and larger in large fields (for large $k$ values), whereas the opposite is true for the USA. Thus, Chinese scholars make more (fewer) contributions to large (cold) subfields compared with their average contribution to all subfields. These results further indicate that Chinese scientists are more likely to follow large topics. We have not investigated potential institutional reasons for this qualitative difference in behavior between Chinese and US scholars, although we strongly suspect that causes may exist at a systematic level. In our definition of $R(k)$, in principle, the numerator $\sum \mu m_k^\alpha$ should eliminate the influence of the absolute sizes of the fields. Therefore, proportionally speaking, in the absence of any other determining factor, the relative contribution ratio should be approximately constant, as in the case of mathematics in Germany and Japan, as illustrated in Fig. 5(e). It should be interesting to investigate the possible reasons behind the observed unevenness of the relative contributions to these fields. An educated guess is that the early development of a research program may be regarded as a catch-up period, during which it is easier to focus more strongly on hotter topics, resulting in smaller contributions in cold subfields and larger contributions in large subfields. If this is the case, then it appears that in
the field of economics, research in Germany, Japan and China is much less developed compared with research in the USA. This is partially confirmed by the small percentage contributions from Germany, Japan and China.

With respect to the numbers of authors, affiliations and references: The more the stronger, with exceptions

Next, we investigate the influence of the numbers of authors, affiliations and references on the degree of tracking large topics. We classify the papers into subsets according to the numbers of authors, affiliations and references and calculate the largeness-tracing exponent $\alpha$ for each subset. It has been found that in recent years, the size of teams and the share of multi-university collaborations have been larger than ever, and the work of such teams tends to attract more citations (Adams, 2012). In addition, it has also been claimed that research works with many authors or many affiliations are typically more focused on large topics (Wei et al., 2013, Adams, 2012). We would like to examine the relation between team size and the degree of largeness tracing. Knowing whether work from large teams is more likely to be largeness-driven rather than interest- or value-driven may have policy value with regard to explicit support for the building of large teams.

As shown in Fig. 6(a), $\alpha$, the exponent quantifying the tracing of large topics, increases with an increasing number of authors, from 0.63 to 0.82 in physics, from 0.87 to 1.2 in mathematics and from 0.55 to 0.68 in economics. These results indicate that large teams typically focus more on large topics, which is consistent with the argument that studies with many authors are typically more focused on large topics (Adams, 2012). Therefore, on average, establishing large teams is not necessarily a good strategy for encouraging value-driven research. However, the exponent $\alpha$ does not increase with an increasing number of authors in biomedical science. This implies that in biomedical science, large teams are not necessary less interest- or value-driven than small teams. In this field, forming large teams might be a good strategy for encouraging value-driven research. Thus, we see that differences between disciplines lead to differences in good team-building strategies. A pure educated guess will be that in the field of biomedical science, there are a significant amount of valuable projects which only large teams can do.
Fig. 6(b) shows that the exponent $\alpha$ increases with an increasing number of affiliations $N_{Aff}$, from 0.65 to 0.83 in physics, from 0.92 to 1.23 in mathematics and from 0.6 to 0.72 in economics. Thus, the above results indicate that on average, cross-affiliation collaborations are not necessarily a good strategy for encouraging value-driven research.

As seen from Fig. 6(c), the exponent $\alpha$ also increases with an increasing number of references $N_{Ref}$, from 0.56 to 0.75 in physics and from 0.82 to 1.01 mathematics, indicating that papers that cite a larger number of references are more likely to address large topics. Interestingly, however, in economics, the exponent $\alpha$ decreases overall with an increasing number of references, from 0.73 to 0.44, implying that papers with a larger number of references are more likely to focus on value-driven research topics. The reasons for this exceptional observation regarding the correlation between largeness tracing and the number of references in economics as well as the exceptional correlation between largeness tracing and the number of authors in biomedical science are not yet known to us.

In principle, one can try to adjust $t_0$ to see more recent trends. We have performed the same analysis with various $t_0$ in measuring sizes of topics. Overall, our above observations on the exponent $\alpha$ and $R (k)$ still hold for shorter periods between $t_0$ and $t$, although their numerical values are different. Moreover, we also have performed the same analysis on various level of MeSH code, such as 6-digit, 9-digit and 15-digit. It is found that the results are qualitatively similar with a different exponent.

**Conclusion and Discussion**

In this paper, we investigate the phenomenon of largeness tracing among scientists from four disciplines, namely, physics, mathematics, economics and biomedical science, all of which have yielded a large number of papers with associated subject classification codes. It is found that overall, scientists in all four disciplines are more likely to publish in large fields, but surprisingly, the values of the exponent $\alpha$ are larger for mathematicians than for the others. We also compare the degrees of largeness tracing between top journals and other journals and find that in all four disciplines, papers published in top journals are less largeness-driven than those in other journals.

In addition, we find that Chinese researchers trace largeness more strongly than US, German and Japanese researchers in all these disciplines. Moreover, it seems that large teams and cross affiliation collaborations are not necessarily beneficial for encouraging the pursuit of value-driven studies: We find that on average, papers with larger numbers of authors, affiliations and references tend to exhibit larger exponents of largeness tracing, although there are interesting exceptions. Papers with more authors are more largeness-driven in physics, mathematics and economics, whereas in biomedical science, such papers are, in fact, less largeness-driven. It is found that cross-affiliation collaborations are more likely to choose...
large topics in physics, mathematics and economics. In addition, the degree of largeness tracing increases with an increasing number of references in physics and mathematics, whereas the opposite is true in economics. It is not surprising that papers with large numbers of references are not particularly value-driven or innovative, so this observation in economics is quite non-trivial. The nature of the unique characteristics of biomedical science and economics that cause them to be exceptions to the generally observed trends remains an open and interesting question.

Scientific innovation is an important goal of science. Ideally, we would hope all scientists choose their research topics mainly based on the scientific values of the questions. Of course, it will be even better if scientific value and largeness agree with each other to a large extent. Otherwise, for individual researchers pursuing innovation is risky (Foster et al., 2015) and maybe working on large topics is a better strategy. However, at levels of nations or societies, an important goal of science policy is to encourage scientific innovation. For that, we might need a policy to balance innovation and largeness tracing. For example, we have seen that Chinese scholars trace largeness most severely and we speculate that it is related to the fact that researchers got additional bonus payment according to the total number of published SCI papers and also the fact that large teams and big players got a dominantly large portion of national research funds. Publications of large teams, according to our preliminary results, are rather more largeness tracing. In fact, it should be further examined in our future studies whether or not large teams and big players are becoming more and more dominant in China. If it is indeed so, then it in a way helps to make a better sense of our discovered empirical results.

The results presented in this paper are based on the available data. In particular, the investigated physics papers are limited to all papers published by APS journals, which is only a subset of all physics journals. In principle, we should collect all papers published in physics journals, but many journals do not use PACS numbers. Moreover, in PubMed datasets, the affiliation information is often inaccurate and the references are not included, and at present, we have not merged PubMed data with WoS data. In this work, we regard papers published in top journals as high-impact papers. In principle, however, we should consider other measures of impact, such as normalized measures (Leydesdorff et al., 2013) or measures based on network analysis (Bergstrom et al., 2008). A corresponding study of technological developments could also be interesting, for example, by considering largeness tracing in patents. In this work, we used established hierarchical classification systems to define topics and as we mentioned earlier, clustering methods based on citations might identify different topics and thus lead to different classifications of papers. How various classification methods influence largeness tracing should also be investigated in the future.

References


Using Machine Learning to Identify Novel Awards – NSF Material Awards as a Case Study

Ting Chen¹  Guopeng Li¹  Xiaomei Wang¹

¹ chenting@casipm.ac.cn  ¹ liguopeng@casipm.ac.cn  ¹ wangxm@casipm.ac.cn
Institutes of Science and Development, Chinese Academy of Sciences (China)

Abstract
Most novelty detection studies in Bibliometrics or S&T management are focused on finding academic papers or patents, in this work, we proposed a novelty detection approach for scientific awards based data mining and machine learning algorithms, designed a machine learning model for identifying novel awards from previous awards structure. Application abstracts of previous and recent awards have been used as training data and predict data, then applying text clustering, dimensionality reduction and one-class SVM classifier to identify the novel data from previous training data. A performance evaluation experiment was designed and implemented, the average accuracy of the proposed approach is 73.78%. A case study of six years on the material awards from the National Science Foundation (NSF) was exemplified, 41 potential emerging awards have been detected from previous awards. This approach could help researchers or decision makers to detect possible novel awards at an early stage, furthermore track the newest evolution of a research field from a scientific funding perspective.

Conference Topic
Knowledge discovery and data mining

Introduction
Discovering the emerging or novel research at the early stage is important to researchers and scholars, there have been many studies in this area. Most of the emerging or novelty studies were focused on finding novel papers or topics. Many related works have been done by researchers in the following approaches: such as text mining and data mining approaches (Glänzel & Thijs, 2011; Zitt & Bassecoulard, 2008; Tu & Seng, 2012), literature-related discovery (Kostoff 2008; Chen 2006; Berry et al. 2009), topological analysis of citation networks (Shibata et al. 2007, 2011); And combinations of other methods or other data, for example patent data (Daim et al. 2006; Verhoeven et al. 2016). As we know the published paper reflects the past work of researchers, the novel research paper or topics detected based on published papers maybe already outdated, but scientific awards represent current and even future research focuses, especially the awards funded by National research funding agencies. If there was an approach that could extract the layout of scientific funding and detect the novel researches at an early stage, it could help researchers and decision makers to track trends and predict the direction of scientific development. There have not been many studies on funding information analysis, particularly novelty detection based on funding data. Most studies for funding analysis were based on statistics or related paper analysis (Huang et al. 2006; Liu et al. 2006, 2008; Jacob & Lefgren, 2011).

This study proposed a novelty detection approach for funding information based on data mining and machine learning algorithms that attempts to discover potential “new” awards. The novel awards that were identified from previous awards which have a better opportunity to be more innovative research. This approach would minimize the number of potential awards, helping researchers or decision markers to track the evolution of a research field.
Methods

In this section, we briefly describe the technical background, the algorithms, and procedures that have been applied. The first subsection refers to the outlines of the proposed machine learning approach; This is followed by the description of the data collection, the text clustering algorithm, the Linear discriminant analysis (LDA) for dimensionality reduction and the one-class SVM classifier for novelty detection. The second subsection designed and implemented a performance evaluation experiment. Four test sets have been tested with proposed approach.

Proposed novelty detection approach

The following methods were proposed to find awards that are “different” or have not been seen in previous awards. The “different” or “new” awards may represent a new research content or direction. As can be seen in Figure 1, the suggested approach employs various methods such as text mining feature extraction, text clustering, dimensionality reduction and novelty detection by classification. The One-class SVM classifier was used as a novelty detection method to find novel awards in this research, but due to the complexity of the classifier, dimensionality reduction techniques needed to be applied in the pre-processing step before later classification for improving computational speed. The One-class SVM classifier will detect the soft boundary of previous “normal” awards in a low-dimensional vector space, then classify new awards as similar or different to that set. As the involvement of many methods and complex algorithms may lead to conceptual misunderstanding and imprecise use in practice, the proposed approach is designed to be executed in four steps: data collection and pre-processing; Find structure of previous awards by K-means++ clustering algorithm; Apply linear discriminant analysis (LDA), a supervised dimensionality reduction technique, to reduce features and enhance computational speed. The K-means clustering labels were used for improving the dimensionality reduction effect; The Final step is identifying novel awards from previous awards via the One-class SVM classifier; The proposed approach described in the following step by step.

Step 1: Data collection and pre-processing

Every year, NSF receives approximately 30,000 new or renewal support proposals for research, graduate and postdoctoral fellowships, and approves approximately 9,000 new awards (Liu et al. 2006). In order to accomplish the proposed approach, more than 2900 awards data from 2010 to 2016 were downloaded from the NSF material department. Downloaded data included title, abstract and start date as the most important information. Because the goal of this research is finding new scientific awards, some unnecessary information contained in abstracts, like applicant and college background, non-technical parts and broader future impact parts were removed. Also, any awards purely supporting international conferences or global cooperation were removed too.

Step 2: Find structure of awards by text Clustering

To detect novel awards from a particular area of awards based on application abstracts, we assumed that if all the awards were grouped into sets by similarity, all of the similar awards
should have been distributed around the centroid of each group set, but perhaps novel awards would have been distributed far away from centroid, so the structure of awards needed to be found first. An improved K-means++ clustering algorithm has been chosen to subdivide previous awards into clusters in this paper, which was proposed in 2007 by Arthur David Arthur and Sergei Vassilvitskii, it is an algorithm for choosing the initial values (or "seeds") for the k-means clustering algorithm to avoid the sometimes poor clustering found by the standard k-means algorithm. The performance of k-means++ clustering algorithm with NSF awards data has been proven by our previous work. (Chen et al. 2015) But some as the standard k-means algorithm, to find the right number of clusters is a challenging issue in clustering analysis, for which no unique solution exists (Zoubi & Rawi, 2008). A method was devised to combine the means of silhouette coefficients and human expert’s option to select the right value of K.

**Step 3: Dimensionality reduction for improving computational speed**
The novel detection model often creates a kernel based model, which means that it is not very scalable. The abstract of awards application usually contains a large number of TF-IDF features, with the purpose of avoiding the overfitting problem in the classification with high dimensional space (Joachims 1998) and reduce computational costs, the high dimensional space needs to project onto a lower-dimensional space with good class-separability. The Linear discriminant analysis (LDA) is a generalization of Fisher's linear discriminant (Fisher 1936), it is the most commonly used as dimensionality reduction technique in pre-processing step before later classification. LDA is very similar to Principal Component Analysis (PCA), they are both inner transformation techniques that are commonly used for dimensionality reduction, but LAD is “supervised” and computes the directions that will represent the axes that maximize the separation between multiple classes. For novelty detection based on structure, the K-means clustering labels can be used as training labels in reducing dimensions, the LDA dimensionality reduction technique is more suitable in this case. The comparison of the two-dimensional projection of awards text features with LDA and PCA is shown in Figure 2, LDA had much better class-separability than PCA with awards data.

![Figure 2: Comparison of 2D projection of LDA and PCA with awards data](image)

**Step 4: Identify emerging awards via One-class SVM**
Novelty detection is the identification of new or unknown data that a machine learning system has not been trained with and was not previously aware of (Markou & Singh, 2003), it is the fundamental requirement of a good classification system. A machine learning system can never have been trained on all possible object classes whose data the system is likely to encounter hence the performance of the network will be poor for those classes that are under-represented in the training set. In this study, the one-class SVM classification algorithm with non-linear kernel was used as a novelty detection method to detect new awards that have not been seen in previous awards (training data). It is a type of one-class classification algorithm, also known as unary classification, tries to identify objects of a specific class amongst all objects, by
learning from a training set containing only the objects of that class (David 2001). This module is particularly useful in scenarios where you have a lot of "normal" data and not many cases of the anomalies you are trying to detect. In this case, the low dimensional LDA features of the past four year’s awards were used as training data, the One-class SVM model will detect the soft boundary of that set in a low-dimensional vector space, in other words, the model will draw a circle in vector space to delineate the scope of “normal” awards, if any new awards were distributed outside the scope, the model will classify it as a novel award.

Performance evaluation
To test the performance of purposed novelty detection approach, an evaluation experiment designed based on NSF materials awards. 1914 NSF material awards in a period of 2012-2015 were used as test data for the experiment. The flowchart of the performance evaluation experiment is shown in Figure 4.

1. Construct dataset with truth ground
To test the performance of the classifier in this awards data, we need to construct a test dataset with truth ground, because it is no labelled novelty award within this research dataset, so the truth ground dataset must be created manually. Firstly, the K-means++ clustering test datasets were applied, which found 30 clusters from the datasets and used clustering labels as truth ground labels for later evaluation tests. Then each cluster was split into previous awards (2012-2014) and recent awards (2015) by a range of year, randomly picking 6 clusters from previous awards as training data and using the training data to train the novelty detection model later. After training data was created, we needed to make test data for the model. The test data should include some awards that are similar to the training data, some awards are different from the training data, so we picked any of the four clusters from the recent data that are the same as the training data, then added the awards from two random clusters that differed from the training data. The awards from the two different clusters were considered as novel awards.

2. Test performance
Use the training data to train the One-class SVM to obtain the machine learning model, then apply the test data to the model for predicting novel awards, awards from two different clusters should be predicted as novel awards, then compare predicted result with truth label and calculate the accuracy of trained model prediction. The result of experiment as shown in table 1, tested the proposed approach four times with different random training and predict data sets, the average accuracy of the model prediction is 73.78%, the accuracy is not particularly high because the truth ground labels are not known for awards data, no human expert involved to identify the real emerging awards during this research. The clustering labels used as truth ground for this were not real truth ground for our test data, so our results are acceptable
### Table 1. Performance evaluation experiments results

<table>
<thead>
<tr>
<th>No. of experiment</th>
<th>Training clusters</th>
<th>Internal predict clusters</th>
<th>External predict clusters</th>
<th>Predict Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6, 20, 8, 0, 25, 24</td>
<td>8, 25, 6, 0</td>
<td>22, 29</td>
<td>76.8%</td>
</tr>
<tr>
<td>2</td>
<td>10, 16, 24, 25, 2, 25</td>
<td>2,10,24,25</td>
<td>7,18</td>
<td>72.3%</td>
</tr>
<tr>
<td>3</td>
<td>7, 18, 1, 30, 11, 17</td>
<td>7, 1, 17, 18</td>
<td>29,15</td>
<td>75.2%</td>
</tr>
<tr>
<td>4</td>
<td>3, 25, 7, 22, 7, 2</td>
<td>25, 7, 3, 22</td>
<td>1, 18</td>
<td>79.8%</td>
</tr>
</tbody>
</table>

### Case Study

A case study of novelty detection for NSF material awards is presented. In order to find out the novel awards these material awards, 2967 NSF material awards collected from the year 2010 to 2016, 2818 research awards remained as case study data after cleaning processing, previous awards data was used to train the novelty detection model which was mentioned earlier, the trained one-class SVM model was then used to predict recent potential emerging awards from training data structure, detailed as follows: Step 1: The documents have been split into training data and predict data by range of year. The training data contained 2362 awards which being awards starting from 2010 to 2014, predict data contained 456 awards which being awards starting from 2015 to 2016. Step 2: Convert training data to TF-IDF vector space which contains a total 6439 words, then subdivided it into 17 clusters through K-means++ clustering algorithms with means of silhouette coefficients and expert judgment. Clustering results and NLP extracted keywords of each cluster are shown in the appendix table 4. Predict data was also converted into the same TF-IDF vector space and used it as feathers of predict data. Step 3: After clustering, both label information and document TF-IDF features were used to reduce dimensionality via LDA. In this research, 6439 training TF-IDF features were reduced to 100 LDA features with good class separation. Step 4: LDA features of training data were used for training the one-class SVM model, which then determined novel data from test data by learning from training data, the parameters of one-class SVM model used in this case study has shown in table 2.

### Table 2: Parameters of one-class SVM mode

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Nu</th>
<th>Gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td>rbf</td>
<td>0.035</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Novelty detection model predicted 41 novel awards (456 awards in predict data) from 2362 training awards which listed in table 3, 14 clusters have novel awards, but 3 other clusters (cluster 2,7,8) do not have any novel awards. Two-dimensional projection of learning results, shown in Figure 4, the dots are awards data in this case study, the white dots are training data (previous awards), the red lines are the learned frontiers from training data, the green dots are predicted data (recent awards) which are considered normal awards that are distributed within the learned frontiers, and the red dots are also predict data, but are considered novel data that is distributed outside of the frontiers.

According to the well distributed two-dimensional projection of training awards and predict awards seen in Figure 4, the results of the proposed model for this case study are reliable. Firstly, the clustering approach applied to the NSF awards provided good clustering results. Each cluster had a clear silhouette showing good separation. Secondly, most novel awards detected
by the learning model have distributed far away from the frontier line of each cluster, this shows
the novel awards detected by the model are quite different from previous awards, those awards
have a good possibility to be identified as emerging awards.

Table 3: 2015-2016 NSF Material Novel Awards

<table>
<thead>
<tr>
<th>Award Title</th>
<th>Cls id</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controlling and quantifying two-level systems, disorder and ideality in tetrahedrally bonded amorphous thin films</td>
<td>0</td>
</tr>
<tr>
<td>Elastic and inelastic scattering studies of supercooled metallic glass-forming liquids - the connection between ordering and fragility</td>
<td>0</td>
</tr>
<tr>
<td>GOALI: Collaborative Research: Understanding Composition-Structure-Chemical Durability Relationships in Multicomponent Oxide Glasses: Influence of Mixed Network Former Effect Structure and Dynamics of Metallic Glass Alloys</td>
<td>0</td>
</tr>
<tr>
<td>MRI: Acquisition of an Analytical Transmission Electron Microscope for High-resolution, Rapid Nanoscale Compositional Mapping of Earth, Planetary, and Advanced Materials</td>
<td>1</td>
</tr>
<tr>
<td>CAREER: Quantum Photonics with Quantum Dots in van der Waals Heterostructures</td>
<td>3</td>
</tr>
<tr>
<td>New Photophysical Processes in Impurity Doped Quantum Dots</td>
<td>3</td>
</tr>
<tr>
<td>Quantum from Classical: Creation of Quantum States of Motion in Nanomechanical Resonators</td>
<td>3</td>
</tr>
<tr>
<td>Charge Effects on Optoelectronic Properties of Nanorod Heterostructures</td>
<td>4</td>
</tr>
<tr>
<td>Collaborative Research: RESOURCE AND REPOSITORY: BROADER IMPACTS OF THE NSF-CMP PROGRAM</td>
<td>4</td>
</tr>
<tr>
<td>Lateral Epitaxial Growth of Nanowires for Electronics</td>
<td>4</td>
</tr>
<tr>
<td>Resource and Repository: Broader Impacts of the NSF-CMP Program</td>
<td>4</td>
</tr>
<tr>
<td>Nonequilibrium States of Topological Quantum Fluids and Unconventional Superconductors</td>
<td>5</td>
</tr>
<tr>
<td>Topological Insulators by Band-Gap Engineering</td>
<td>5</td>
</tr>
<tr>
<td>CAREER: Supramolecular Polycondensation of Polymer Brushes via DNA Hybridization</td>
<td>6</td>
</tr>
<tr>
<td>Charge transport and trap-healing effect at semiconductor/polymer heterointerfaces.</td>
<td>9</td>
</tr>
<tr>
<td>Heterostructures and Superlattices of Two-Dimensional Layered Materials</td>
<td>9</td>
</tr>
<tr>
<td>Manipulating 2D Superconductivity through atomic scale control of boundary conditions</td>
<td>9</td>
</tr>
<tr>
<td>Probing and manipulating strained interfaces with oxide superconductors</td>
<td>9</td>
</tr>
<tr>
<td>Studies of Mixed Polymer Brushes Designed for Periodic In-Plane Order</td>
<td>10</td>
</tr>
<tr>
<td>Investigation of Strongly Correlated Itinerant Magnets and Potential Quantum Spin Liquids</td>
<td>11</td>
</tr>
<tr>
<td>MRI: Acquisition of a High Field, Wide Temperature Range Electrical, Magnetic and Thermal Properties Measurement System</td>
<td>11</td>
</tr>
<tr>
<td>New Directions for Organic Spintronics: Organic-Based Magnetic Heterostructures and Microwave Magnetodynamics</td>
<td>11</td>
</tr>
<tr>
<td>CAREER: Control of Charge Carrier Dynamics in Complex Thermoelectric Semiconductors</td>
<td>12</td>
</tr>
<tr>
<td>Colloidal Nanocrystal Routes to Inorganic Nanocomposite Thermoelectric Materials</td>
<td>12</td>
</tr>
<tr>
<td>DMREF: Collaborative Research: Extreme Bandgap Semiconductors</td>
<td>12</td>
</tr>
<tr>
<td>Combined Macroscopic and Nanoscopic Studies of the Photovoltaic Behavior of Organic Perovskite Materials</td>
<td>13</td>
</tr>
<tr>
<td>Femtosecond Microscopy of Charge Transport in Perovskite Thin Films</td>
<td>13</td>
</tr>
<tr>
<td>Tuning the Photovoltaic Performance of Benzothiophene and Benzodifuran Polymeric Semiconductors by Molecular Design</td>
<td>13</td>
</tr>
<tr>
<td>CAREER:Designing Hund's Metals from Transition Metal Sulfdies</td>
<td>14</td>
</tr>
<tr>
<td>Integrated, Functional MOF@Polymer Membranes</td>
<td>14</td>
</tr>
<tr>
<td>Screw Dislocation-Driven Growth of Complex Nanomaterials</td>
<td>14</td>
</tr>
<tr>
<td>Synthetic Integration of Inorganic-Organic Functionality in Hybrid Semiconductors Based on Chalcogenide Clusters</td>
<td>14</td>
</tr>
<tr>
<td>CAREER: Two-dimensional Van Der Waals systems with tunable spin-orbit coupling</td>
<td>15</td>
</tr>
<tr>
<td>MRI: Development of a Mid-infrared Optical Microscope for Investigation of Femtosecond Dynamics of Single Large Spin Orbit Semiconductor Heterostructures</td>
<td>15</td>
</tr>
<tr>
<td>Physics of Strong Disorder and Correlation</td>
<td>15</td>
</tr>
<tr>
<td>Realizing and Manipulating Magnetism and Transport in Two-Dimensional Transition Metal Dichalcogenides</td>
<td>15</td>
</tr>
<tr>
<td>Electronic and Magnetic Phenomena in Iron-based Superconductors</td>
<td>16</td>
</tr>
<tr>
<td>Interplay Between Superconductivity and Charge Order in Near-Critical Metals</td>
<td>16</td>
</tr>
<tr>
<td>RUI: Investigation of Strongly Correlated Electron Behavior in Rare Earth Related Materials</td>
<td>16</td>
</tr>
</tbody>
</table>
Conclusion and limitations

This study has proposed a machine learning approach to discover potential novel awards based on funding application documents. Firstly, we designed an approach combining a clustering algorithm, dimensionality reduction technique and one-class SVM algorithm to identify recent novel awards from previous awards. Secondly, a performance evaluation experiment was designed and implemented using clustering predicted labels as truth ground labels to test the performance of the proposed approach (average of 73.78% accuracy). Finally, regarding practical implementation, this study adopted a case study based on NSF material awards from 2010 to 2016, a total of 2363 awards, where 41 potential emerging awards have been detected. As this study is only at the explorative stage, it is subject to certain limitations as outlined here. First, data cleaning of application documents is not fully automated, a lot of manual work was involved in this step. Second, this approach detects novel awards based on an awards structure, awards should be subdivided into groups first. The accuracy of novelty detection is dependent on clustering performance, also the One SVM classifier is one of the most commonly used novelty detection algorithm, but may not be the best one, we will test more algorithm and find the most the most accurate one in future, also to obtain good clustering performance required participation of human experts. Finally, the results of the case study are also not verified by human experts due to the vast variety of NSF material awards and the difficulty of judgment for novel awards, hopefully, in future human experts can be involved in furthering this research to improve the accuracy of the approach.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos: 71173211)

References


Appendix

Table 3: 15-16 Training Awards Clustering Results

<table>
<thead>
<tr>
<th>Cluster id</th>
<th>Size</th>
<th>Cluster NLP keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>51</td>
<td>metallic glasses, Bulk Metallic Glasses, Develop Metallic Glasses, Multicomponent Oxide Glasses, Ion-conducting Chalcogenide Glasses, Engineering Organic Glasses, multicomponent borosilicate glasses, Exceptionally Stable Glasses</td>
</tr>
</tbody>
</table>

320
alloys, High Entropy Alloys, Nanocrystalline Alloys, Semiconductor Alloys, Non-Cubic Metal Alloys,
Ti-base Nanocrystalline Alloys, Nanocrystalline Fe-Cr-Ni Alloys, Supersaturated Structural Alloys
Complex Oxide Heterostructures, Epitaxial Oxide Heterostructures, molecular beam epitaxy, Oxide
Semiconductor Heterostructures, metal oxide, Metal Oxides, Oxide Epitaxial Growth, Metal Oxide
Surfaces, Complex Metal Oxides
polymer, Conjugated Polymer, conjugated polymers, Polymer Films, Block Copolymers, Oriented
Conjugated Polymer, Polymer Crystallization, Dispersed Polymer Melts, Functional Polymer,
Nanostructured Conducting Polymer
magnetic nanostructures, high magnetic fields, high magnetic field, Novel Magnetic Materials,
Perpendicular Magnetic Anisotropy, Nanoscale Magnetic Interactions, Magnetic Bubble Dynamics,
Ferromagnetic Shape Memory
Thermoelectric Materials, Thermoelectric Properties, Oxide Thermoelectric Materials, Complex
Thermoelectric Semiconductors, Thermoelectric Nanocomposite Materials, Nanocomposite
Thermoelectric Materials
solar cells, Solar Energy Conversion, Polymer Solar Cells, Organic Solar Cells, Nanostructured Solar
Cells, hybrid solar cells, Solar Energy Harvesting, solar collaborative, Based Solar Cells, InGaN Solar
Cells, Organic Photovoltaic Materials
liquid crystals, Crystal Growth, molecular crystals, Smectic Liquid Crystals, liquid crystalline materials,
Single Crystals, metal-organic frameworks, Electronic Liquid Crystals, Porous Molecular Crystals,
Organic Molecular Crystal
spin, Spin Transport, spin-orbit coupling, Low-Dimensional Spin Systems, Ballistic Spin Transport,
Spin Dynamics, Single Spin Tunneling, Electron Spin Dynamics, Spin Orbit Systems, Cooperative Spin
Clusters, Spin Injection
strongly correlated electron, correlated electron systems, superconductors, Novel Superconductors, iron-
based superconductors, Pnictide Superconductors, iron-based high-temperature superconductors
Measuring Scientific Knowledge Flows by Deploying Citation Context Analysis using Machine Learning Approach on PLoS ONE Full Text

Saeed-Ul Hassan¹, Anam Akram¹, Awais Asghar¹, Naif Radi Aljohani²

¹ saeed-ul-hassan@itu.edu.pk
Information Technology University, Lahore (Pakistan)

² nraljohani@kau.edu.sa
King Abdulaziz University, Jeddah (Kingdom of Saudi Arabia)

Abstract
We measure the knowledge flows among the countries by analyzing publication and citation data. We argue that not all citations are equally important, therefore, in contrast to existing techniques that utilize absolute citation counts to quantify knowledge flows among different entities, our model employs citation context analysis technique using machine learning approach to distinguish between important and non-important citations. We use 14 novel features (including context based, cue words based and textual based) to train Random Forest classifier on labeled dataset of 20,527 publications downloaded from the Association for Computational Linguistics anthology (http://allenai.org/data.html). Finally, we present a case study to elucidate our deployed method on the dataset of PLoS ONE full text publications in the field of Computer and Information Sciences. Our results show significant volume of important knowledge flows from the United States consumed by the international scientific community. Of the total knowledge flow from China, we find relatively less proportion (only 4.11%) fall under the category of important knowledge flow. We believe that such analyses are helpful to understand the dynamics of the relevant knowledge flows across the nations.

Conference Topic
Knowledge discovery and data mining

Introduction
In the past few years, the world has seen remarkable development of scientific knowledge-research interactions at progressively quick rate (Schott, 1994). Researchers have tried to represent knowledge as static and have developed various methods to extract, discover and reason about it. However, knowledge is dynamic and requires human brain for its discovery, circulation, synthesis, generalization, and problem solving. Scholarly publications are the most significant medium that transfers knowledge between scientific communities (Yang & Wang, 2015; Zhuge, 2006). In addition to measure knowledge flows among the scientists – this phenomenon has been measured at institutional or country level (Ingwersen et. at, 2000; Albert & Adam, 2003). Undoubtedly, knowledge has become crucially important in global economies, since many economies are becoming knowledge based. The emerging economies are participating in the production of new scientific knowledge that contribute to an accelerated pace of technical and scientific advance. The rapid growth of scientific publications is one key indicator to quantify the production of these new knowledge powerhouses. In last two decades, the world has seen a remarkable growth in the production of scientific knowledge. While the scientific literature index in Scopus database has increased 2.27 times in 2014 compared with ~1.14 million publications in 1996, China has remarkably published 15.8 times more in 2014 compared with ~28k publications in 1996. This significant increase in creating new knowledge sources is a key indicator to measure the scientific advancement of knowledge driven economies. However, the retentions of these knowledge sources does not
solely be contingent on single isolated entity, rather depends upon transfer and exchange among the knowledge based economies (Hassan & Haddawy, 2015a; Hassan & Haddawy, 2015b). Recently, knowledge flows have been measured by analyzing citation networks among the publications produced by the scholars from different parts of the world (Hassan & Haddawy, 2015a; Hassan & Haddawy, 2015b). Citations indeed provide a quantitative way to measure the knowledge flows among the published works but not all citations are equal. While some certainly indicate that the cited work is conceptually used or, more importantly, extended in the new publication, some are incidental, e.g., they discuss the cited work in the context of related work that does not directly impact the new effort. Therefore, it is important to differentiate between non-important (incidental) and important (conceptual) citations because measuring knowledge flow using absolute citations counts would not be adequate.

This paper measures knowledge flows among the countries by analyzing publications and citations data. Unlike existing techniques that utilize absolute citation counts to quantify knowledge flows among the entities, our proposed model utilizes citation context to distinguish important knowledge flows. The followings are the objectives of this paper: a) to study the knowledge flows using citation exchange among the papers published by the scientific communities across the countries; b) to identify the ratio of important knowledge flows w.r.t to total knowledge flows from the countries; 3) and to study the global network of knowledge flows across the countries. The remaining paper is organized as follows. The related work is discussed in the next section followed by our deployed methodology and data collection process. Further, we present a case study to elucidate our deployed method in the field of Computer Science. Lastly, we present concluding remarks along with future directions.

**Literature review**

This section presents a brief literature review in two directions: while the first half discusses related studies that deploy citation analysis to measure knowledge across different entities, the second half presents a review on methods on classification of citations in order to distinguish most important knowledge flows.

**Knowledge Flows among the scientific literature**

Indeed, knowledge does not exist in isolation; rather, it is defused, circulated and exchanged among various entities. Citations among the papers have been employed to measure knowledge flows in science (Zhuge, 2006). By aggregating citations among the paper into several higher levels, scientists have presented journal level, university level, scientific discipline level and national level knowledge flows by analyzing citation exchange of the following entities: journal/conference citations (Borgman & Rice, 1992; Leydesdorff and Probst, 2009; Zhao & Wu, 2014), institution citations (Börner et. al, 2006; Yan & Sugimoto, 2011, Liu et al., 2014), scientific discipline citations (Van Leeuwen & Tijssen, 2000; Yan et al., 2013) and country citations (Hassan & Haddawy, 2013).

At the journal level, examining citation exchange is used to study patterns across the fields. Leydesdorff and Probst (2009) point knowledge out flow of political science and social psychology journals on communication science journals. At the institute level, a study of spatio-temporal changes of 500 most cited institutes shows an inverse relation between the log of distance and log of citations among the institutes (Börner et. al, 2006). This results in favor that knowledge flows are less dependent on physical distances (Yan & Sugimoto, 2011, Liu et al., 2014). At field level, publications by Cronin & Davenport (1989) and Cronin & Pearson (1990) may consider pioneers to study knowledge exchange among different disciplines. Their work deploys trading terminologies and considers only two ends of knowledge flows ecosystem i.e. knowledge exporter and knowledge importer but lack to discuss concrete knowledge flow paths. However, recent work of Yan (2015, 2016), explores knowledge paths among scientific
disciplines. At country level, Hassan & Haddawy (2013) introduce a quantitative approach to measure the international scholarly impact and present geographical maps to visualize knowledge flows across the countries. Zhang et al (2013) analyze the database of the entire publication of the American Physical Society and discovered the trends of production and consumption in physics in different cities.

Apart from the quantitative representations of knowledge flows, scientists have been working on semantic representations for better understanding of knowledge flows. Zhuge (2006) is one of the pioneers who proposed several techniques for the discovery of knowledge flows in science by deploying hyper cycle model and semantic link network in e-science environment. He also presented an effective e-learning system using semantic link network to find out semantic relationship between learning resources and communities to support effective browsing of knowledge (Zhuge, 2009; 2010; Zhuge, 2011). Furthermore, Zhang et. al. (2013) analyzed the database of the entire publication of the American Physical Society and discovered the trends of production and consumption in physics in different cities. More recently, a methodology to semantically analyze the knowledge flow across countries using Latent Dirichlet Allocation model and Euclidian distance was proposed by Hassan and Haddawy (2015a, 2015b). Their method consisted of identifying the research topic of the publications produced by a given country in a given research area. Then they determine the scholarly impact of the actual research topics in the cited topics across the countries.

Citation Classification Methods

According to Borgman(1990) and Luukkonen (1992), citations are commonly used to measure the impact of a research article by how often they are cited, however, researchers have always criticized the pure quantitative analysis of citations quoting that most citations are done out of “politeness, policy or piety” (Ziman, 1968). It has been argued that citations used in negation or passing citations should not weigh as much as central citations in a paper (Bonzi, 1982). One of the best-known and most widely used schemes for citation functions is devised by Moravscik & Murugesan (1975). They divide citations into four categories namely conceptual vs operational, juxtapositional or evolutionary, perfunctory or organic and finally negational or confirmative. Chubin and Moitra (1975) slightly modify this scheme by making each of the categories mutually exclusive. Oppenheim & Renn (1978) devise a unique scheme to classify citations in the physical sciences to determine why old papers are cited more. Small (1986) devise a simple classification scheme for cancer virology citations using cue words. Garzone (1997) developed an automated classification mechanism for citations using the linguistic semantic grammar. Teufel et al. (2006) also devise a supervised learning model to classify citation function utilizing linguistically-inspired features. Furthermore, the work on the proximity of co-cited references by Liu & Chen (2011) while conducting co-citation analysis and the work of Hu et al (2013) that present a model to learn from the citation instances distributed in 350 full text scientific articles and are also notable ones.

More recently, Valenzuela et al (2015) introduce a novel task of identifying important citations in scholarly literature, i.e., citations that indicate that the cited work is used or extended in the citing document. They devise a mechanism to separate important and meaningful citations from the unimportant ones. They model this task as a supervised classification problem at two levels of detail: a coarse one which classes important vs. non-important and a more detailed one which classifies into four categories i.e. Incidental (related work), Incidental (comparison), Important (using the work) and Important (extending the work). Hassan et al (2017) extend this work by introducing six novel features that outperforms the results produced by Valenzuela et al (2015) with producing very promising 0.91 ROC area under the curve with Random Forest classifier on the same dataset.
Summary

The presented literature review discusses solid theoretical and methodological foundation for quantitative and semantic representations of knowledge flows among several entities: scientific papers, scholars, institutions, disciplines and regions using citations analysis. In addition, we also present an overview of citations classification methods that we use to distinguish important relevant knowledge flows across the entities. We observe that the literature review of knowledge flows and citations classifications is discussed in two separate research streams. The goal of this research is to fill this gap by presenting holistic view of knowledge flows in order to distinguish most important one, since using simple absolute citations counts is not adequate.

Data and Methodology

Data

We measure scientific knowledge flows across the countries using publications and citations on PLoS ONE dataset. We download 4,138 papers in XML format indexed in Computer and Information Sciences field during Jan 01, 2014 to Sep 01, 2015. Since the cited references do not provide affiliation related information of authors, therefore, we procure metadata information such as (affiliation, abstract, author defined keywords etc.) of 31,839 unique references of our downloaded dataset from Scopus database using its API. Note that we find 21,804 matches in Scopus data since not all references in PLoS ONE are indexed in Scopus. This final set of 4,138 papers and 21,804 references is used to measures scientific knowledge flows across the countries. Table 1 shows country wise distribution of PLoS ONE publications for the countries having at least 2% publication in our dataset. Note that we use double counting method that gives full credit to publications co-authored by the institutions from different countries. In addition, Table 2 shows country wise distribution of references indexed in Scopus database cited in the selected PLoS ONE publications. Since the Scopus API only provides affiliation information of first author, therefore, in reference dataset we assign each publication to only one country corresponding to the affiliation of first author. Thus, for the reference dataset the implication of double counting method is not applicable.

Table 1: Country wise distribution of selected dataset during Jan 01, 2014 to Sep 01, 2015.

<table>
<thead>
<tr>
<th>Country</th>
<th>*P</th>
<th>(P/**T_P)X100</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>1209</td>
<td>23%</td>
</tr>
<tr>
<td>China</td>
<td>660</td>
<td>12%</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>440</td>
<td>8%</td>
</tr>
<tr>
<td>Germany</td>
<td>341</td>
<td>6%</td>
</tr>
<tr>
<td>Canada</td>
<td>193</td>
<td>4%</td>
</tr>
<tr>
<td>France</td>
<td>180</td>
<td>3%</td>
</tr>
<tr>
<td>Australia</td>
<td>175</td>
<td>3%</td>
</tr>
<tr>
<td>Netherlands</td>
<td>174</td>
<td>3%</td>
</tr>
<tr>
<td>Spain</td>
<td>168</td>
<td>3%</td>
</tr>
<tr>
<td>Italy</td>
<td>159</td>
<td>3%</td>
</tr>
<tr>
<td>Brazil</td>
<td>132</td>
<td>2%</td>
</tr>
<tr>
<td>Japan</td>
<td>131</td>
<td>2%</td>
</tr>
<tr>
<td>Switzerland</td>
<td>98</td>
<td>2%</td>
</tr>
<tr>
<td>Sweden</td>
<td>83</td>
<td>2%</td>
</tr>
</tbody>
</table>

*P = Publications by the authors affiliated with an institution in a given country; **T_P = Total publications downloaded from PLoS ONE.
Table 2: Country wise distribution of references (indexed in Scopus) cited in selected PLoS ONE publications dataset.

<table>
<thead>
<tr>
<th>Country</th>
<th>*P</th>
<th>(P/**T_R)*X100</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>15941</td>
<td>34%</td>
</tr>
<tr>
<td>China</td>
<td>4286</td>
<td>9%</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>4037</td>
<td>9%</td>
</tr>
<tr>
<td>Germany</td>
<td>3346</td>
<td>7%</td>
</tr>
<tr>
<td>Canada</td>
<td>1848</td>
<td>4%</td>
</tr>
<tr>
<td>France</td>
<td>1627</td>
<td>3%</td>
</tr>
<tr>
<td>Australia</td>
<td>1509</td>
<td>3%</td>
</tr>
<tr>
<td>Netherlands</td>
<td>1479</td>
<td>3%</td>
</tr>
<tr>
<td>Spain</td>
<td>1193</td>
<td>3%</td>
</tr>
<tr>
<td>Japan</td>
<td>1182</td>
<td>3%</td>
</tr>
<tr>
<td>Italy</td>
<td>1079</td>
<td>2%</td>
</tr>
<tr>
<td>Switzerland</td>
<td>861</td>
<td>2%</td>
</tr>
<tr>
<td>Brazil</td>
<td>812</td>
<td>2%</td>
</tr>
<tr>
<td>Sweden</td>
<td>778</td>
<td>2%</td>
</tr>
</tbody>
</table>

*R= References published by the authors affiliated with an institution in a given country; **T_R= Total references indexed in Scopus citing publications downloaded from PLoS ONE.

Approach

Our machine learning model classifies citations found in full text publications into important and non-important, using Random Forest classifier. The deployed model has three aspects: 1) a given cited paper is important to a citing paper if it is cited several times in full text; 2) the location of a given citation in full text is important to determine its influence on the citing paper e.g. citation in method sections would probably be more important than the citation in related work; 3) and the text around the citation in full text since the it would indicate the reason for the citation. Using the above aspects, we selected total 14 novel features (eight context based, four cue words based and two textual based) to classify cited references into important and non-important.

Features for classification

The followings are context based feathers: F1 = Total number of citations received by a reference till Dec 2016; F2 = Number of citations from the current paper to the cited paper; F3 = Citations in introduction section; F4 = Citations in literature review section; F5 = Citations in method section; F6 = Citations in experiment section; F7 = Citations in discussion section; F8 = Citations in conclusion section. The cue word based feathers are: F9 = Cue words for Related work citations; F10 = Cue words for Comparative citations; F11 = Cue words for Using the existing work; F12 = Cue words for Extending the existing work. Finally, the textual based features are: F13 = Similarity between the abstract of cited paper and text of citing paper and F14 = Author overlap - we set it to TRUE if the cited paper and citing paper share at least one common author.

Experimental setup

Finally, to prepare data for experiments, we normalize all features by centering on the mean and scaling to unit variance. We generate 46,956 number of records in our feature data which naturally exceed from the total number of unique references. Since we find many papers in...
PLoS ONE dataset that share a given reference. In addition, a given reference may also be cited more than ones in a given paper. Therefore, for each pair of cited reference and citing paper (or context within a paper) a separate record is inserted in feature dataset. For learning, we implement Random Forest classifier in python using the Scikit Learn toolkit that run all features with equal weights. We used Gini index in Random Forest as a criterion to measure the quality of a split. For the experiment, we set the max_feature to auto that allow the classifier to consider all the features that seem optimal in every tree because our dataset consist of 14 features that does not need any selection restrictions.

For training the model, we use labeled dataset of 20,527 publications from the Association for Computational Linguistics anthology (http://allenai.org/data.html). Figure 1 presents the pseudo code of the method deployed.

<table>
<thead>
<tr>
<th>Pseudo Code</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Require:</strong></td>
</tr>
<tr>
<td>1. Training data D</td>
</tr>
<tr>
<td>2. Test data T</td>
</tr>
<tr>
<td>3. Feature set</td>
</tr>
</tbody>
</table>

**Pre-process and feature extraction:**

4. Web links → download corpus
5. XMLtoText in Python→ Convert XML files to plain text
6. Python Code ← Input plain text files
7. Fetch citation text and occurrence ← Using Python
8. Tokenize ← Citation text
9. Remove punctuation, numbers, HTML tags stop words ← Using Python
10. Lemmatization ← Python
11. POS tagger → Get synonyms using wordnet in R
12. Extract features
   a. Total citation count received by reference ← using Google scholar
   i. Normalize count by (2016-y)
   b. Occurrence of citation to cited paper ← Using Python
   c. Section titles → Normalize titles → count total citations in each section
   d. Cue words for each class, citation text → Cosine similarity
   e. Citation text, Reference abstract → cosine similarity
   f. Authors list of cited and citing paper → Return true if any match

**Classification procedure:**

13. Normalize features
14. Train on the labeled dataset D
15. Make prediction for each instance in test dataset T and obtain results

**Figure 1: Pseudo code of deployed methodology**

**Results and discussion**

In this section, we show the results of our case study that presents the knowledge flows across the nations using all citations, important citations, their ratios. We also show comprehensive citation networks across the countries to identify nations that exhibit highest knowledge flow impact across the glob using important citations.
Figure 2: Distribution of references published by the authors across the countries cited in the papers published by the authors from the United States

Figure 2 shows the distribution of references published by the authors across the countries to the papers published by the authors from the United States. Of the total references appear in the publication by the United States, 8692 references (53.24%) are from the papers published by the United States. This shows that a large proportion of references are self cited - if we treat the whole country as a unit. In other words, more than half of the knowledge produced by the United States is consumed within the country. Despite the fact, China’s huge volume of publication output, of the total references appear in the United States publications only 5.21% are from China in our dataset. Interestingly, the dependence of the United States on the knowledge produced by its own scientific community is even more significant when it comes to the most important one. Of the total important references cited by the authors from the United States in PLoS ONE publication dataset, about 75% (738 references) are authored by the authors from the United States. This shows huge influence of important knowledge produced by the United States within the country. Interestingly, the significance of knowledge flow from China is even reduced when it comes to important knowledge flows only. This shows even weaker dependence of the United States on the knowledge produced by the Chinese scientific community.

Figure 3: Distribution of references published by the authors across the countries cited in the papers published by the authors from the United Kingdom

Figure 3 shows the distribution of references published by the authors across the countries to the papers published by the authors from the United Kingdom. Of the total references appear in the publication by the United States, 1636 references (27.90%) are from the papers published by the United States, followed by 26.02% from the United Kingdom. This shows that more than one fourth of the total knowledge consumed by the United Kingdom comes from the United States. However, the United Kingdom shows 47.81% (186 references) of self consumption of important knowledge flows.
Figure 4 shows the references appear in the publications by the Canadian scientific community. Of the total, 912 references (i.e. 30.55%) are from the papers published by the United States, followed by 23.85% from Canada. In terms of important knowledge flow towards Canada, we find increase in the the significance of knowledge produced by the United States. Similarly, the United States appears to be the main contributor whose knowledge is consumed by France, Australia and Japan.

In contrast, the analyses of the distribution of references cited in the papers published by China shows that unlike many other nations in our dataset, the United States is not the main contributor of knowledge that is consumed by China (see Fig. 5). We show that of the total references appear in the publications by China, 2934 references (i.e. 39.2%) are from the papers published by China itself. Interestingly, the dependence of China on the knowledge produced by its own scientific community is even more significant when it comes to the most important flows with more than 40% (130 references) are authored by the authors from China. The case of China in terms of self dependence on its own knowledge resources is somewhat similar to the United States. Interestingly, we find similar pattern of knowledge self dependence in case of Germany and the Netherlands in our dataset.

Figure 5: Distribution of references published by the authors across the countries cited in the papers published by the authors from China

In addition, to analyze the distribution of knowledge consumed by a given country from other countries, we also show the proportion of important knowledge flow from a given country w.r.t to total knowledge flow from that country. Table 3 shows the list of countries having more than 30 important references in our dataset. We find Austria appears at the top in the list with 13.71% of total published references fall under the category of important ones. The Netherlands and Germany appear among the top five despite a relatively large volume of total references published by these nations. Interestingly, Spain, Japan, France and China show less than 5% of their total published references classified as important.
Table 3: Country wise proportion of important references w.r.t total references (indexed in Scopus) cited in selected PLoS ONE publications dataset.

<table>
<thead>
<tr>
<th>Country</th>
<th>*R</th>
<th>*IMP_R</th>
<th>(IMP_R/R) x100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>372</td>
<td>51</td>
<td>13.71%</td>
</tr>
<tr>
<td>Sweden</td>
<td>778</td>
<td>76</td>
<td>9.77%</td>
</tr>
<tr>
<td>Brazil</td>
<td>812</td>
<td>74</td>
<td>9.11%</td>
</tr>
<tr>
<td>Netherlands</td>
<td>1479</td>
<td>134</td>
<td>9.06%</td>
</tr>
<tr>
<td>Germany</td>
<td>3346</td>
<td>246</td>
<td>7.35%</td>
</tr>
<tr>
<td>Taiwan</td>
<td>442</td>
<td>31</td>
<td>7.01%</td>
</tr>
<tr>
<td>United States</td>
<td>15941</td>
<td>1089</td>
<td>6.83%</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>4037</td>
<td>250</td>
<td>6.19%</td>
</tr>
<tr>
<td>Australia</td>
<td>1509</td>
<td>84</td>
<td>5.57%</td>
</tr>
<tr>
<td>Canada</td>
<td>1848</td>
<td>96</td>
<td>5.19%</td>
</tr>
<tr>
<td>Switzerland</td>
<td>861</td>
<td>44</td>
<td>5.11%</td>
</tr>
<tr>
<td>Spain</td>
<td>1193</td>
<td>57</td>
<td>4.78%</td>
</tr>
<tr>
<td>Japan</td>
<td>1182</td>
<td>55</td>
<td>4.65%</td>
</tr>
<tr>
<td>France</td>
<td>1627</td>
<td>72</td>
<td>4.43%</td>
</tr>
<tr>
<td>China</td>
<td>4286</td>
<td>176</td>
<td>4.11%</td>
</tr>
</tbody>
</table>

*R= References published by the authors affiliated with an institution in a given country; ** IMP_R= Important references published by the authors affiliated with an institution in a given country.

Figure 6: Knowledge flow network (using all references) across the countries
Finally, by deploying the mapping technique proposed by Rosvall & Bergstrom (2008), we draw two networks to analyze the knowledge flows across the countries. We show 22 nodes and 145 direct links for each network, where a node represents a given country and the flow from a given country $A$ to county $B$ represents the references published by $A$ and cited in the paper published by $B$. Similarly, the flow from a given country $B$ to county $A$ represents the references published by $B$ and cited in the paper published by $A$. Here, color intensity, thickness of the directed line and size of a node shows the volume of knowledge flow. While the inner circle in a given node shows the knowledge flow within the country, the outer circle shows knowledge flow outside the country.

Figure 6 demonstrates the knowledge flows across the countries. We find the United States, followed by China, the United Kingdom, Germany and the Netherlands appear to be the most influential nodes in the network. We find significant knowledge flow from the United States to other nations. Further, analyzing only the important knowledge flows reveals even more influential role of the United States across the countries (see Fig. 7). Interestingly, of the total important knowledge flow from China, a large proportion is consumed within the country. Overall, the significance of the United States is raised in the network of important flows (Fig. 7) compared to the network created by using both important and non-important ones (Fig. 6).

![Knowledge flow network]  
*Figure 7: Knowledge flow network (using important references) across the countries*
Concluding remarks

In this paper, we have measured the knowledge flows among the countries by analyzing publication and citation data. Our deployed Random Forest based machine learning model classifies important and non-important citations which we have further used to distinguish between important and non-important knowledge flows. Our results have shown that despite an exponential increase in publication output by the Chinese scientific community, the most important global knowledge flows that are consumed by the international scientific community are still exhibited by the United States, followed by the United Kingdom and Germany. We have also highlighted that in contrast to many other nations analyzed in our dataset, the dependence of China on the knowledge produced by its own scientific community is even more significant when it comes to the most important knowledge flows produced by China.

In summation, the machine learning approach deployed in this paper helps to distinguish between important knowledge flows. In addition, the method deployed is also fundamental for the systems that detect and track emerging research themes and to measure the impact of scientific literature in increasingly scholarly big data. In future, using the deployed methodology, we plan to build bibliometric enhanced information systems for improved retrieval of bibliography databases. Finally, we also plan to improve our machine learning model by including more features such as citation sentiments (negative, positive or neutral) - we believe that this would significantly improve the accuracy of our model.

References


Effect of publication month on citation impact

Abstract
Bibliometric studies most commonly take whole years as temporal units of analysis, implicitly treating all publications as if they were published at the exact same point in time. This leads to systematic bias in favor of early-months publications and to the detriment of late-months publications in citation analyses. This contribution demonstrates the size of this distortion on a large body of publications from all disciplines over citation windows of up to 15 years using bivariate correlations between publication month and two citation impact indicators. It is found that early-month publications enjoy a substantial citation advantage which arises from citations of the first three years after publication and remains detectable for much longer citation windows.

Conference Topic
Data Accuracy and disambiguation; Methods and techniques

Introduction
Citation impact normalization is a core methodological concept for the construction of advanced bibliometric indicators, i.e. those which remove the effects of different scientific discipline, the kind of document and the time of publication. By delineating sets of publications similar to each other with regards to content and formal characteristics and using these sets to compute reference values and relative impact indicators, heterogeneity due to these factors is reduced. The intention is to make fair comparisons; that is, to compare like with like (Schubert and Braun, 1986). The basic formal characteristics, as opposed to the content (disciplinary area), that are taken into account are publication type (such as journal papers, conference proceedings papers, books and book chapters), document type (such as research articles, review articles, letters, editorials etc.) and publication date.

One important component of normalization is controlling for publication time, as, ceteris paribus, the more time has passed since publication the more citations will accrue. The publication year is commonly used to define publication periods. This practice is based on the implicit assumption that, for the question of interest of a study, it makes no difference when exactly in a year a paper was published. The fact that documents published in January have eleven months more to be read and cited than works published in December of the same year raises the question if this assumption is justified, and, if it is not always justified, under which circumstances and how a more precise publication date should be taken into account in citation analysis.

The question of the influence of a more exact publication date is also related to the problem of finding adequate citation windows, the time period in which citations to papers in a set of publications are counted. A too short citation window, say two years, would clearly lead to bias against papers published towards the end of the investigation period compared to those published towards the beginning. Consider the following simple illustration. Citations are tallied at the end of the year after publication (2-year citation window). Then papers from January had 23 months to be read and cited while December papers had 11 months less, just 12 months, which is 52 % of the time period of the January papers. This relative disadvantage becomes smaller as the citation window length is increased. In a five year citation window, for example, the December papers had 83 % of the citation time of the January papers.

Recently, Gai et al. (2015) studied the month citation bias in one field, one specialty and one journal for citation windows from 1 to 9 years. They exclude a few very highly cited papers and group publication months into six two-month periods. For papers in physical geography they find significant group differences for citation counts for up to five years after publication. The results are similar for papers on the topic of diabetes. For the Journal of
Biological Chemistry, group differences are significant up to the seventh year after publication.

As that study been limited to specific disciplines I investigate if the effect is general to all disciplines, as far as covered in Web of Science journal collections. Two disciplines with very different citation dynamics are investigated additionally to provide insight into field differences. Field differences in publication month effect on citation counts can be hypothesized to mirror fields’ size (in terms of number of papers) and citation speed. The present study takes into account citation windows of up to 15 years. Month effects are measured for the two bibliometric indicators mean citations per paper and share of highly cited papers for single-months groups of papers. The online publication date is tested for influence on citations in addition to that of the issue publication date. Finally, it is demonstrated how to eliminate publication month bias by constructing citation windows of equal length for month cohorts of papers.

Data and methods

Datasets
Dataset A consists of all journal publications of document type ‘Article’ from the year 2000, obtained from Web of Science \( (n = 767,959) \), for which publication month data was either available in the source data or could be estimated. 6539 articles from 2000 had to be excluded from analysis because they had neither a publication date value nor enough contextual data to obtain estimates of publication month. The number of citations to the articles in each year were computed from citation link data constructed using the ‘iFQ’ procedure described in Olensky, Schmidt & van Eck (2016) which has improved recall compared to WoS-provided citation links, using references data from all journal publications that appeared between 2000 and 2014. Citations from identical citing to cited first authors were removed. It was found that neither including all self-citations or excluding all self-citations affected results substantially.

Dataset B will be used to assess the combined influence of online and print publication months. It is comprised of articles contained in PubMed, based on 2016 base files. Publications of the print publication year 2009 which had a DOI datum available were selected and matched to WoS records based on the DOI. The records were restricted to the WoS document types ‘Article’ and ‘Review’. This results in 253,292 publications. Citation counts were computed by CWTS by their method which was also described in Olensky, Schmidt & van Eck (2016). Self-citations are not removed.

Indicators and statistical methods
The indicators of citation impact for which differences in the values of month-cohorts of publications are investigated are the average citations per paper and the share of highly cited papers. In the case of the latter, the share of papers of each publication month among the 10% most highly cited papers over all months. The statistical association of the impact indicators with the months variable is measured with the Pearson correlation coefficient on the level of monthly aggregates of publications. In the case of the association with individual papers’ citation counts, the correlation is also computed for the log-transformed citation count in order to minimize the effect of the non-normality of the citation distribution. To assess the compound effects of issue and online publication month on individual publication citation counts, we perform linear regression with log-transformed citation counts as the dependent variable (Thelwall, 2016).
Results

Effect of publication month on citations at the level of month-cohort groups of papers

To examine the effect of the publication date on citations to papers published within one year, all publications published in each month are considered as natural groups. The average numbers of citations per paper (CPP) of all month groups are calculated in order to characterize the groups’ citation impact. I examine first the average CPP per month based on citations from single years for the first four years. The results are shown in Table 1. For the first three years the earlier months’ publications receive on average more citations. In years one and two the relationship between publication month and CPP is almost exactly linear negative. The association weakens in year three to \( r = -0.74 \) and in year 4 the effect is no longer present.

Table 1. Mean CPP by month for individual years

<table>
<thead>
<tr>
<th>month</th>
<th>mean year 1 citations</th>
<th>mean year 2 citations</th>
<th>mean year 3 citations</th>
<th>mean year 4 citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.73</td>
<td>1.91</td>
<td>2.20</td>
<td>2.19</td>
</tr>
<tr>
<td>2</td>
<td>0.54</td>
<td>1.59</td>
<td>1.91</td>
<td>1.93</td>
</tr>
<tr>
<td>3</td>
<td>0.51</td>
<td>1.73</td>
<td>2.09</td>
<td>2.11</td>
</tr>
<tr>
<td>4</td>
<td>0.41</td>
<td>1.59</td>
<td>2.02</td>
<td>2.05</td>
</tr>
<tr>
<td>5</td>
<td>0.31</td>
<td>1.49</td>
<td>1.93</td>
<td>1.97</td>
</tr>
<tr>
<td>6</td>
<td>0.25</td>
<td>1.36</td>
<td>1.86</td>
<td>1.93</td>
</tr>
<tr>
<td>7</td>
<td>0.21</td>
<td>1.51</td>
<td>2.10</td>
<td>2.16</td>
</tr>
<tr>
<td>8</td>
<td>0.14</td>
<td>1.21</td>
<td>1.84</td>
<td>1.93</td>
</tr>
<tr>
<td>9</td>
<td>0.12</td>
<td>1.21</td>
<td>1.96</td>
<td>2.07</td>
</tr>
<tr>
<td>10</td>
<td>0.08</td>
<td>1.04</td>
<td>1.83</td>
<td>2.00</td>
</tr>
<tr>
<td>11</td>
<td>0.07</td>
<td>0.92</td>
<td>1.75</td>
<td>1.92</td>
</tr>
<tr>
<td>12</td>
<td>0.05</td>
<td>0.84</td>
<td>1.79</td>
<td>2.00</td>
</tr>
</tbody>
</table>

Pearson r

\[-0.96^* \quad -0.97^* \quad -0.74^* \quad -0.37\]

Note: * \( p < 0.01 \)

For year one, the year of publication, December papers have almost no time at all to be cited. In year two, the CPP of December papers is 44% that of the January papers’. For year three December papers reach on average 82% of the CPP of January papers and for year four 82% as well. However, at that point the January publications are no longer the most highly cited and the December papers are not the least cited any more.

Clearly, early-month publications enjoy a considerable citation advantage for three years. Having established the effect for single citation years, we now move to citation windows of multiple years. Citation windows of two and three years after publication will clearly exhibit the correlation. However, how long does it take for the effect of those early years to become negligible? In order to answer this question in the next step of analysis citations are cumulated for subsequent years.

The results presented in Table 2 indicate that the effect decreases rather slowly. The correlation is as high as \(-0.8\) after nine years, \(-0.7\) after eleven years and at the end of the observation period, after 15 years, it is down to \(-0.6\). The correlation remains statistically significant at the 1% level up to year 11.

It could be possible that CPP is affected more by publication month than other impact indicators because citation count distributions are very skewed and the mean of this distribution is less robust than indicators based on other characteristics of the distribution. The means of the citation counts of the month cohorts might be distorted in this way. To rule out such an effect, the publication month bias is also assessed with another impact indicator, the
share of highly cited publications. In a publication set not influenced by a publication month effect, about equal shares of highly cited publications are expected be present in every month group. We consider the share of each month’s publications among the approximately 10% most highly cited publications. The issue of the publications having a citation count exactly equal to the threshold value, the 0.9-quantile, is disregarded here, as it of no consequence with respect to the central issue of this analysis, namely the association of publication month and indicator values. The computed values for selected citation windows, presented in Table 3, show the results of the calculations. In that table, the row labeled “actual share of publications >= threshold” shows the exact percentage of articles that have as much or more citations than the threshold value, which is given one row above. In the absence of bias, each month’s percentage should be very close to that overall percentage, which is the expected value. For example, in the column for year 1 to 8 (9-year citation window) 11.8% of January papers reach or exceed the threshold citation count while 9.1% of December papers do. Evidently, citation impact as operationalized by this indicator is also affected by publication month bias, even up to very long citation windows.
### Table 2. Mean CPP by month for selected cumulative citation windows

<table>
<thead>
<tr>
<th>Month</th>
<th>1 to 4</th>
<th>1 to 5</th>
<th>1 to 6</th>
<th>1 to 8</th>
<th>1 to 10</th>
<th>1 to 12</th>
<th>1 to 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.04</td>
<td>9.18</td>
<td>11.25</td>
<td>15.20</td>
<td>18.88</td>
<td>22.29</td>
<td>25.54</td>
</tr>
<tr>
<td>2</td>
<td>5.98</td>
<td>7.89</td>
<td>9.73</td>
<td>13.29</td>
<td>16.66</td>
<td>19.83</td>
<td>22.85</td>
</tr>
<tr>
<td>3</td>
<td>6.44</td>
<td>8.53</td>
<td>10.54</td>
<td>14.41</td>
<td>18.04</td>
<td>21.41</td>
<td>24.61</td>
</tr>
<tr>
<td>4</td>
<td>6.06</td>
<td>8.08</td>
<td>10.04</td>
<td>13.79</td>
<td>17.32</td>
<td>20.58</td>
<td>23.70</td>
</tr>
<tr>
<td>5</td>
<td>5.70</td>
<td>7.66</td>
<td>9.55</td>
<td>13.19</td>
<td>16.62</td>
<td>19.82</td>
<td>22.89</td>
</tr>
<tr>
<td>6</td>
<td>5.40</td>
<td>7.32</td>
<td>9.19</td>
<td>12.80</td>
<td>16.21</td>
<td>19.42</td>
<td>22.49</td>
</tr>
<tr>
<td>7</td>
<td>5.98</td>
<td>8.12</td>
<td>10.20</td>
<td>14.17</td>
<td>17.90</td>
<td>21.33</td>
<td>24.60</td>
</tr>
<tr>
<td>8</td>
<td>5.12</td>
<td>7.05</td>
<td>8.95</td>
<td>12.60</td>
<td>16.08</td>
<td>19.31</td>
<td>22.42</td>
</tr>
<tr>
<td>9</td>
<td>5.35</td>
<td>7.43</td>
<td>9.45</td>
<td>13.35</td>
<td>17.02</td>
<td>20.42</td>
<td>23.68</td>
</tr>
<tr>
<td>10</td>
<td>4.95</td>
<td>6.97</td>
<td>8.95</td>
<td>12.77</td>
<td>16.37</td>
<td>19.72</td>
<td>22.92</td>
</tr>
<tr>
<td>11</td>
<td>4.66</td>
<td>6.60</td>
<td>8.50</td>
<td>12.18</td>
<td>15.65</td>
<td>18.88</td>
<td>21.96</td>
</tr>
<tr>
<td>12</td>
<td>4.69</td>
<td>6.73</td>
<td>8.73</td>
<td>12.59</td>
<td>16.24</td>
<td>19.62</td>
<td>22.85</td>
</tr>
</tbody>
</table>

Pearson r: -0.92, -0.88, -0.83, -0.75, -0.68, -0.63, -0.57
p-value: <0.01, <0.01, <0.01, <0.01, 0.01, 0.02, 0.03

Note: correlation: Pearson r of month (1 to 12) and average CPP in a citation window

### Table 3. Share of highly cited publications by month for selected cumulative citation windows

<table>
<thead>
<tr>
<th>Month</th>
<th>1 to 4</th>
<th>1 to 5</th>
<th>1 to 6</th>
<th>1 to 8</th>
<th>1 to 10</th>
<th>1 to 12</th>
<th>1 to 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.69</td>
<td>12.63</td>
<td>12.72</td>
<td>11.95</td>
<td>11.70</td>
<td>11.32</td>
<td>11.20</td>
</tr>
<tr>
<td>3</td>
<td>12.29</td>
<td>11.64</td>
<td>11.76</td>
<td>11.30</td>
<td>11.16</td>
<td>10.85</td>
<td>10.79</td>
</tr>
<tr>
<td>4</td>
<td>11.43</td>
<td>10.91</td>
<td>11.15</td>
<td>11.74</td>
<td>10.74</td>
<td>10.57</td>
<td>10.32</td>
</tr>
<tr>
<td>5</td>
<td>10.72</td>
<td>10.26</td>
<td>10.54</td>
<td>10.24</td>
<td>10.16</td>
<td>9.98</td>
<td>10.06</td>
</tr>
<tr>
<td>7</td>
<td>11.13</td>
<td>10.73</td>
<td>11.16</td>
<td>10.94</td>
<td>10.97</td>
<td>10.73</td>
<td>10.82</td>
</tr>
<tr>
<td>9</td>
<td>9.74</td>
<td>9.69</td>
<td>10.16</td>
<td>10.10</td>
<td>10.25</td>
<td>10.16</td>
<td>10.21</td>
</tr>
<tr>
<td>10</td>
<td>8.61</td>
<td>8.86</td>
<td>9.50</td>
<td>9.63</td>
<td>9.84</td>
<td>9.79</td>
<td>9.93</td>
</tr>
<tr>
<td>12</td>
<td>7.93</td>
<td>8.36</td>
<td>9.08</td>
<td>9.33</td>
<td>9.60</td>
<td>9.53</td>
<td>9.69</td>
</tr>
</tbody>
</table>

Pearson r: -0.93, -0.90, -0.86, -0.80, -0.72, -0.66, -0.60
p-value: <0.01, <0.01, <0.01, <0.01, <0.01, <0.05, <0.05

* 95 % confidence interval includes 0
Effect of publication month on the level of individual publications

I turn now to the association between citation scores of individual articles and publication month for the entire dataset A and the two discipline subsets that can be formed by selecting only papers in the WoS subject categories of Cell Biology (N=14510) and Mathematics (N=5538). Contrary to the averages and ratios used earlier for month-based groups of papers, the distribution of individual article citation counts exhibits severe skewness. Therefore Table 4 shows correlations for both untransformed and log-transformed citation counts for selected citation windows. Pearson correlations are intended to measure the linear associations between normally distributed variables. The log-transformation of citation count plus 1 results in a distribution approximating a Gaussian. The associations measured by Pearson correlation coefficient are small but statistically significant. The reason for this is the very large variability of citation counts due to reasons other than publication month. Nevertheless, the explanatory power of publication month is considerable on the individual article level shortly after publication. The results are quite similar for the entire dataset and the mathematics and cell biology subsets.

<table>
<thead>
<tr>
<th>citation window and transformation</th>
<th>3y cits</th>
<th>log(1+(3y cits))</th>
<th>5y cits</th>
<th>log(1+(5y cits))</th>
<th>10y cits</th>
<th>log(1+(10y cits))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson r entire dataset</td>
<td>-0.066</td>
<td>-0.100</td>
<td>-0.036</td>
<td>-0.052</td>
<td>-0.016</td>
<td>-0.020</td>
</tr>
<tr>
<td>Pearson r cell biology subset</td>
<td>-0.105</td>
<td>-0.105</td>
<td>-0.036</td>
<td>-0.046</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Pearson r mathematics subset</td>
<td>-0.099</td>
<td>-0.108</td>
<td>-0.062</td>
<td>-0.073</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

Note: reported correlations have p-values less than 0.01.

“3/5/10y cits” is short for 3/5/10 year cumulative citation counts

Effect of online publication month

Up to this point, the publication month of the completed issue has been studied. Individual articles might be published online, usually in advance, at a date quite different from the publication date of the issue. This raises the question if it might not be better to use the online publication date instead of the issue date to eliminate publication date bias. Dataset B (sourced from PubMed) is used to study this issue.

Print publication month and online publication month are counted as months since January of year 1 CE, so that for example January 2009 has the value 24109. The correlation between the two variables is 0.77. The Pearson correlation of print publication month and citation count on the article level for a three-year citation window for this dataset is -0.081 and the correlation of month and log(1+3y cits) it is -0.144. These associations are a little stronger than for the cell biology subset of dataset A. The correlation of online publication month and three-year citations is 0.007 and that of online publication month and log-transformed citation count is 0.040. For all five values, p<0.01 holds.

We use linear regression to assess the individual and combined relationship of print and online publication months on papers’ citation counts. The independent variable is the natural logarithm of citations after three years plus one, in order to account for the heavy skewness of the citation distribution (variable cits3y), following the recommendation in Thelwall (2016). The transformation results in a variable with mean=1.79, sd=0.95 and skewness=0.13. Three models for different specifications of the independent variables are used to investigate the individual and combined influence of relative earlier publication in print and online of papers from the same year on citation counts. The three dependent variables are pub.month (print
issue publication month), *epub.month* (online publication date month) and *abs.d* (absolute difference of *pub.month* and *e-pub.month*). Specifications (1) and (2) in Table 5 confirm that both print issue month and online publication date month are individually significant predictors of citations count, being responsible for a 4% reduction and a 1% reduction in 3-year citation count on the original scale per increase in one month. Specification (3) includes the variable *abs.d* in addition to issue publication month. This leads to an improvement in total predictive power compared to either variable. These findings suggest that it is useful to include both the print issue and the difference between issue and online date as control variables in regression analyses of citation count data. The two publication date variables are complementary to some degree and it is thus not advisable to substitute one for the other.

**Table 5. Regression of 3-year citations on print and online publication months**

<table>
<thead>
<tr>
<th>specification</th>
<th>Dependent variable: log(1 + cits3y)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td><em>pub.month</em></td>
<td>-0.040* (0.001)</td>
</tr>
<tr>
<td><em>epub.month</em></td>
<td>-0.009* (0.0004)</td>
</tr>
<tr>
<td><em>abs.d</em></td>
<td>-0.043* (0.001)</td>
</tr>
<tr>
<td><em>R²</em></td>
<td>0.021 0.002 0.036</td>
</tr>
<tr>
<td><em>F Statistic</em></td>
<td>5,386.323* 395.647* 4,690.368*</td>
</tr>
<tr>
<td></td>
<td>(df = 1; df = 1; df = 2;</td>
</tr>
<tr>
<td></td>
<td>253290) 253290) 253289)</td>
</tr>
</tbody>
</table>

**Note:** *p<0.01

**Month level precision citation windows**

It has been shown that publications of different month-cohorts of the same year have systematically shifted citation distributions. It was also found that the effect can be detected on the level of individual publications. The consequences for research using regression methods are as simple as including the publication month and, if possible, the difference of online and issue publication month, as control variables. This approach is not expedient in bibliometrics-informed research assessment. One obvious possibility to eliminate the distortion is to use month-based citation windows. This requires publication month data for both the cited and the citing documents. This allows the calculation of the exact number of citations to source publications within a specific number of months after publication which is equal for all publications, thus eliminating the bias. To verify this proposal, the publication month data was constructed for all citing documents in dataset A. New citation counts based on month-precise citation windows were calculated for 12, 24, 36, etc. months. The citation windows for each month cohort are then of the same length, not of lengths differing up to 11 months. The month-based citation windows are also longer than the corresponding year-based windows. Take for example the 3-year citation window and the 36-month citation window. The three year window ranges from 35 months (from January, year 1, to December, year 3) to 24 months (from December, year 1, to December, year 3). All month-based citation windows are 36 months long. Table 6 shows the average citation counts for some selected citation windows which may be contrasted with the simple year-based method results in Table 1 and
Table 2. In the former case, there is no correlation between publication month and mean CPP. Figure 1 shows the differences in average citation counts of each month’s publications for the two citation window methods in the 3-year case.

<table>
<thead>
<tr>
<th></th>
<th>average cumulative citations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>month 1 year 3 years 5 years 15 years</td>
</tr>
<tr>
<td>1</td>
<td>0.82 5.44 10.31 30.38</td>
</tr>
<tr>
<td>2</td>
<td>0.73 4.83 10.26 28.09</td>
</tr>
<tr>
<td>3</td>
<td>0.82 5.28 10.03 29.79</td>
</tr>
<tr>
<td>4</td>
<td>0.81 5.11 9.72 28.89</td>
</tr>
<tr>
<td>5</td>
<td>0.80 5.00 9.52 28.35</td>
</tr>
<tr>
<td>6</td>
<td>0.78 4.83 9.22 27.67</td>
</tr>
<tr>
<td>7</td>
<td>0.95 5.54 10.35 29.76</td>
</tr>
<tr>
<td>8</td>
<td>0.82 4.96 9.39 27.85</td>
</tr>
<tr>
<td>9</td>
<td>0.90 5.28 9.99 29.07</td>
</tr>
<tr>
<td>10</td>
<td>0.86 5.09 9.68 28.35</td>
</tr>
<tr>
<td>11</td>
<td>0.87 5.04 9.49 27.50</td>
</tr>
<tr>
<td>12</td>
<td>0.86 5.11 9.68 27.96</td>
</tr>
</tbody>
</table>

When month-level precision citations are available, it also is possible to assess how many citations were not included when using year-level precision citations even though they cite an article within the indicated time period. Consider the example of a nominal three-year citation window. In our dataset A, counting citations in the first three years gives a total of 2,740,173 citations. Setting up 36-month citation windows for each publication month cohort, however, gives a total of 3,432,353 citations. The nominal year-based method therefore takes into account 80% of the 36-months citations. Similarly, in the case of a nominal 5-year vs. actual 60-month citation window, the year-based method takes into account 89% of the month-precise method’s citations.
Discussion

It has been demonstrated that publications of early months have a citation advantage compared to publications of late months from the same publication year. This advantage is of considerable magnitude and cannot be eliminated by removing author self-citations. The effect is present in citations from the first three years after publication but influences metrics using cumulative citation windows for many years afterwards.

The present study has some limitations. The study considered only two publication years. No temporal dynamics in publication month bias were investigated. The predictor variable of interest was the point of publication of individual articles within a year which was approximated by using the recorded or estimated month of publication of an issue. The publication month was estimated in a portion of the data with some bias and inaccuracy. Another limitation is that the official month of publication must not necessarily be close to the time when the article could first be read. Since the observed early-month bias is quite large, it should be a matter of concern in particular in situations when distinct sets of publications are compared for their citation impact and the sets are so small as to not allow enough variability in publication months so that the systematic effect may cause distortion. A set of publications will probably not be affected by this bias if it can be demonstrated to be distributed uniformly over months and the number of observations is large enough.

In regression analyses which use citation counts as the dependent variable, publication time point within a year and if available also the absolute value of the difference between print publication date and online publication date in months should routinely be included as control variables.
Acknowledgements
I would like to thank Nees Jan van Eck for providing dataset B and much helpful discussion and suggestions related to this study.

References
Do under-cited influential sleeping beauties exist?

Xiaojun Hu¹Ronald Rousseau²

¹ xjhu@zju.edu.cn
Medical Information Center, Zhejiang University School of Medicine, Hangzhou 310058 (China)

² ronald.rousseau@kuleuven.be
KU Leuven, B-3000 Leuven (Belgium) & University of Antwerp (UA), Faculty of Social Sciences, B-2000 Antwerp, (Belgium)

Abstract
In this study we investigate if sleeping beauties, also known as articles suffering delayed recognition, can at the same time be under-cited influential articles. Theoretically these two types of articles are independent, in the sense that being a sleeping beauty depends on the number and time distribution of received citations, while being an under-cited influential article depends only partially on the number of received (first generation) citations, and much more on second and third citation generations. Among 49 sleeping beauties we found 13 that are also under-cited influential.

Conference Topic
Citation Analysis

Introduction
In diachronous studies and depending on their citation curves or ageing profiles articles can be categorized in different ways. In (Smith Aversa, 1985) the author makes, for highly cited papers, a distinction between delayed rise in received citations combined with a slow decline, on the one hand, and early rise combined with a rapid decline, on the other. More recently, Costas et al. (2010) classify documents into three general types: delayed documents, which receive the main part of their citations later than normal documents; flashes-in-the-pan (van Dalen & Henkens, 2005), which receive citations immediately after their publication but are not cited or influential in the long term; and normal documents, documents with a typical distribution of citations over time. The delayed documents include articles known as sleeping beauties (van Raan, 2004), articles suffering from delayed or late recognition, being premature discoveries, being ahead of one’s time, suffering from Mendel’s syndrome or being late bloomers (all synonyms or near-synonyms, describing the same phenomenon).

Recently, and especially in China, sleeping beauties attracted a lot of attention (Ke et al., 2015; Li, 2014; Li & Shi, 2016; Li & Ye, 2016; Du & Wu, 2016), leading among others to “all-elements-sleeping-beauties” (Li & Ye, 2012), i.e. articles that at first receive some attention, then go through a period in which they are hardly noticed, followed by a period of being highly cited, awakened by a prince, so to speak. Figure 1 shows the citation curve of (Romans, 1986), a typical sleeping beauty studied by van Raan (2004). It received no citations for nine years and then received more than 100 citations over the next six years. After a peak (it received 32 citations in the year 1999) its citation curve shows a gradual and irregular decline similar to that of ‘normally cited’ articles.
A different type of articles: under-cited influential articles

All the types of articles mentioned above were based on the – direct – citation distribution. Yet, recently Hu and Rousseau (2016, 2017) introduced another type of articles in which first, second and third generation citations play a role, leading to the sparkling index (Hu & Rousseau, 2017). These articles are referred to as under-cited influential articles. Such articles are characterized by three properties: (1) they are reasonably well-cited (a basic requirement to be influential); (2) citations of citations (second generation citations) are rather high, so that the original one is influential in an indirect way (a more refined token of influence); (3) given condition two, these articles received fewer citations than expected (being under-cited). Each of these three requirements is operationalized leading to a two formulae referred to as sparkling indices. There are two types of sparking indices depending on the number of received citations: a sparking index on the 1% level, denoted as S1 (for articles that are cited at least 201 times) and a sparking index on the 10% level, denoted as S10 (for articles that are cited at least 21 times but less than 201). Details are provided in (Hu & Rousseau, 2016, 2017). These two articles also show that such under-cited influential articles are neither common, nor extremely rare. Many articles written by Nobel Prize winners, but also other, including recent articles, fall into the group of under-cited influential articles.

Before continuing our investigation we want to clarify the terminology of under-cited influential articles. It is clear that these articles are influential: they are reasonably well cited and, through citing articles, their influence on the development of the ideas that started with them is even higher. Many of them, like the Nobel prize winning articles studied in (Hu & Rousseau, 2017), can even be described as foundational articles. These articles are also referred to as under-cited. By this we do not mean that they are treated unfairly. We just want to stress that many articles that built on them received more citations. One reason could be that these articles added new results that incorporated the results of the ‘under-cited’ one, so that later investigators do not feel the need to refer to the original.

Clearly sleeping beauties and under-cited influential articles differ in many aspects except that in short-term evaluations based on bibliometric indicators, they both are under-valued. Sleeping beauties usually begin with a long period of few citations and then direct citations (first generation) go up. Under-cited influential articles usually are rather well-cited but
The number of received citations does not fully reflect their influence on their field. Probably only expert peer reviewers can see their importance. These considerations led us to the following research question.

**Are there sleeping beauties which are also under-cited influential articles?**

We looked through a set of articles dealing with delayed recognition or sleeping beauties and determined 49 articles considered to be sleeping beauties, restricting then to 1950 or later and included in the SCI-E. As a universally accepted, precise definition of a sleeping beauty does not exist, we do not discuss if these articles are really sleeping beauties. We just accept the opinion of the authors who studied these articles. For all these sleeping beauties we checked if they are also under-cited influential articles. Citations were collected and calculations made in February 2017. Bibliographic information on these sleeping beauties is given in Table 2 (Appendix), while the articles discussing these sleeping beauties can be found in Table 3 (Appendix). Results of our investigations are shown in Table 1. We found 13 sleeping beauties of the 49 checked that are also under-cited influential articles. This is 27%. Of course numbers are too small to accord any meaning to this percentage.

**Table 1. Articles that are sleeping beauties and moreover also under-cited influential**

<table>
<thead>
<tr>
<th>First author</th>
<th>PY</th>
<th>Discussed as sleeping beauty by the following author(s); see Table 3</th>
<th>Received citations</th>
<th>Sparking index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leakey [2]</td>
<td>1964</td>
<td>Tobias</td>
<td>310</td>
<td>S(_1) = 358</td>
</tr>
<tr>
<td>De Rujula [4]</td>
<td>1977</td>
<td>Li</td>
<td>191</td>
<td>S(_{10}) = 210</td>
</tr>
<tr>
<td>Gitter [8]</td>
<td>1980</td>
<td>Glänzel et al.</td>
<td>88</td>
<td>S(_{10}) = 176</td>
</tr>
<tr>
<td>Lis [9]</td>
<td>1980</td>
<td>Glänzel et al.</td>
<td>763</td>
<td>S(_1) = 966</td>
</tr>
<tr>
<td>Ogino [10]</td>
<td>1980</td>
<td>Glänzel et al. /Braun et al.</td>
<td>483</td>
<td>S(_1) = 1078</td>
</tr>
<tr>
<td>Waite [12]</td>
<td>1980</td>
<td>Braun et al.</td>
<td>164</td>
<td>S(_{10}) = 176</td>
</tr>
<tr>
<td>Romans [13]</td>
<td>1986</td>
<td>Van Raan</td>
<td>347</td>
<td>S(_1) = 723</td>
</tr>
</tbody>
</table>

**Characteristics of articles which are sleeping beauties as well as under-cited influential**

The article by Romans (1986) turned out to be also under-cited influential according to our definition. This is illustrated in Figures 2-3 where, besides the original citation curve we also show the curve of the citing curves with the second most citations (in yearly distribution and in cumulative number). As the top 1% (rounded up) citing articles consists of four articles we show the second one among these (referred to as the median citing article). This median citing article, published in 1999, received already in the year 2000 more citations that Romans’ (Figure 2) and the next year its cumulative total citations was higher than Romans’ (Figure 3). This is just an illustration as such an intersection is neither necessary nor sufficient to be an under-cited influential article.
Discussions and conclusions

We note the following caveat: the definitions of sleeping beauties and under-cited influential articles depend on some ad hoc rules. This leads to the question: do these 13 articles, or most of these, really belong to the two types mentioned, or would a slight change in definition to one of the two types make a consequential difference?

Assume that we have a strict definition of a sleeping beauty, then we wonder after how many years is it clear that an article is a sleeping beauty? Is it, at that moment, also possible to say that this sleeping beauty is moreover under-cited influential (for which a strict definition exists), or is it then usually too early, as one needs three generations of citing articles and it took quite some time before the first generation was ‘large enough’ and hence does not easily lead to many second and third generation citations. Assuming that a ‘recent’ sleeping beauty is not under-cited influential is it then possibly to predict (i.e., determine a probability) if it will ever be?
We conclude that we have shown that there do exist articles that are sleeping beauties and are at the same time under-cited influential, Romans’ (1986) being a case in point.

Acknowledgments
This work was supported by the National Natural Science Foundation of China Grant 71573225. The authors would like to thank Zhang Yuning and Hu Xiaoyue for their help in data collection.

References


Li, J. (2014). Citation curves of "all-elements-sleeping-beauties": "flash in the pan" first and then "delayed recognition". *Scientometrics*, 100(2), 595-601.


Appendix
Table 2. List of sleeping beauties studied in this contribution. All articles are published in 1950 or later and are included in the SCI-E. Articles that discuss these sleeping beauties are listed in Table 3.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Title</th>
</tr>
</thead>
</table>
Table 3. Articles discussing sleeping beauties studied in our investigation.


Li, J. (2014). Citation curves of "all-elements-sleeping-beauties": "flash in the pan" first and then "delayed recognition". Scientometrics, 100(2), 595-601.


Aerospace Discipline Study Based on Highly Cited Papers

Qin Ping¹  Xie Ting²   Duan Linbo³

¹ qplib@nuaa.edu.cn
Nanjing University of Aeronautics and Astronautics, Nanjing (China)

² xieting@nuaa.edu.cn
Nanjing University of Aeronautics and Astronautics, Nanjing (China)

³ linboduan@nuaa.edu.cn
Nanjing University of Aeronautics and Astronautics, Nanjing (China)

Abstract
Based on the biblio-metrology methods and the visualization technology, this paper investigates the features of the highly cited papers of the aerospace discipline in Elsevier’s Scopus database, including research hotspots and international cooperation. Through tracking the conferences cited by the highly cited papers, of the highly academic influence in the discipline, the academic journals, conference proceedings, monographs, and the science and technology reports are listed. The research results provide a reference for aerospace discipline study.

Introduction
Established in China in early 20th century, the aerospace science and technology has been developed rapidly in recent years, which is based on mathematics, physics, and the modern technology science, and covers the fields such as aircraft design, propulsion theory and engineering, manufacturing engineering, man-machine and environmental engineering. This paper investigated the highly cited papers and analyzed the important literature characteristics of the discipline via biblio-metrology methods, including the research hotspots, international cooperation, and the academic influence, etc.

Data sources and research methods
Sample data
Scopus database contains over 21,500 titles from more than 5,000 international publishers. Across all research fields—science, mathematics, engineering, technology, health and medicine, social sciences, and arts and humanities—Scopus delivers a broad overview of global, interdisciplinary scientific information that researchers, so that is more suitable for the analysis of high-level papers in the field of engineering. In what follows, we take Aerospace Engineering in Scopus database as example to conduct a comprehensive analysis on highly cited papers published in last five years. By December 2015, the Aerospace Engineering journals have been included 41 kinds of journals issued by 11 countries, including 12 in America, 10 in Britain, 7 in Netherlands, 3 in China, 2 in India, 2 in Brazil, 1 in France, 1 in Russia, 1 in Lithuania, 1 in Germany and 1 in Egypt. Here, a total of 108,415 papers in 2010-2014 were taken in this study.

Definition of highly cited papers
“Highly cited” papers mean that they have world-class influence in the discipline. According to Essential Science indicators (ESI), those papers that the citing frequency accounts for the first 1% are called the “highly cited” papers. In this paper, the top 200 papers of aerospace fields annually are selected as highly cited papers as shown in Table 1, with a total of 1,001 papers within five years.
Table 1. The number of papers of aerospace science within five years

<table>
<thead>
<tr>
<th>Year</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of papers</td>
<td>20,888</td>
<td>19,735</td>
<td>22,559</td>
<td>21,899</td>
<td>23,334</td>
<td>108,415</td>
</tr>
<tr>
<td>Number of highly cited papers</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>201</td>
<td>1,001</td>
</tr>
</tbody>
</table>

### Analysis of Highly Cited Papers

**Paper length**

In general, the length of a paper would have an impact on the quality and the academic level of the paper. By analyzing the lengths of 1,001 highly cited papers, it is found that the page number of the highly cited papers has a broad range from 1 page to 108 pages, with the average length of 13.97 pages, as shown in Table 2. The top-level academic journals in aerospace discipline have no limitation of the length of paper, this gives the authors an enough space to elaborate their research findings.

Table 2. Analysis of the length of highly cited papers

<table>
<thead>
<tr>
<th>Year</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>Average Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page number range</td>
<td>1~44</td>
<td>1~108</td>
<td>1~90</td>
<td>1~41</td>
<td>1~41</td>
<td>1~108</td>
</tr>
<tr>
<td>Median page number</td>
<td>13</td>
<td>14</td>
<td>13</td>
<td>13</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>Average page number</td>
<td>13.53</td>
<td>15.72</td>
<td>13.72</td>
<td>13.75</td>
<td>13.13</td>
<td>13.97</td>
</tr>
</tbody>
</table>

### Analysis of Cited Frequency

The cited frequency is one of the important indicators to evaluate scientific research achievements. From 2010 to 2014, the cited frequency of the highly cited papers in aerospace discipline is in the range of 8 to 317, the average value of which is 40.75, as shown in Table 3. The tendency change of the cited frequency of highly cited papers is shown in Figure 1. It can be seen that the cited frequency of highly cited papers is approximately linearly increasing following the increased age in five years.

Table 3. The cited frequency of highly cited papers

<table>
<thead>
<tr>
<th>Year</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cited frequency (times)</td>
<td>41~300</td>
<td>34~317</td>
<td>28~124</td>
<td>17~121</td>
<td>8~44</td>
<td>8~317</td>
</tr>
<tr>
<td>Median</td>
<td>57</td>
<td>49</td>
<td>35</td>
<td>21</td>
<td>11</td>
<td>36</td>
</tr>
<tr>
<td>Average</td>
<td>67.01</td>
<td>58.47</td>
<td>39.58</td>
<td>25.81</td>
<td>13.10</td>
<td>40.75</td>
</tr>
</tbody>
</table>
Analysis of Publications

Highly cited papers are commonly published in influential publications. For aerospace discipline, 41 journals are cataloged into the influential publications. The first 20 journals published the highly cited papers are listed in Figure 2, in which the IEEE Transactions on Vehicular Technology ranks the first with 220 papers.

International Cooperation Analysis

As an important way, the international cooperation has received increased attention in the recent years, especially in China. In 1001 highly cited papers, most of the published papers are China, the United States, and the United Kingdom. Figure 3 gives the number of highly cited papers of aerospace discipline in the top 17 countries. By using Gephi, the cooperation network between the 17 countries is illustrated as shown in Figure 4, wherein the connection lines with the deeper colour represent much more cooperation between the countries. One can see from...
According to the frequency statistic of the keywords of highly cited papers within five years, it is found that the number of keywords of each highly cited paper ranges from 2 to 16, with an average of 4.8, as shown in Table 4. The frequently used keywords are Cognitive radio, Nanofluid and Fault diagnosis in order. The first 40 high-frequency keywords are shown in Figure 5. The high-frequency keywords can reflect the recent research hotspots of the discipline.

Table 4. The number of keywords of highly cited papers

<table>
<thead>
<tr>
<th>Year</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A total of 41548 references have been cited by 1,001 highly cited papers of aerospace discipline, with the average number of 41.51, listed in Table 5. Notice that the correlation between the number of references and the frequency of cited papers is not obvious. The top three years of most cited references are during 2005-2009, 2010-2014 and 2000-2005, as shown in Figure. 6. Figure 6 shows that the frequency of the cited references is gradually declined with time, with the obvious long-tail effect.

### Table 5. The number of the references of highly cited papers in 2010-2014

<table>
<thead>
<tr>
<th>Year</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>The number of highly cited papers</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>201</td>
<td>1,001</td>
</tr>
<tr>
<td>The number of references</td>
<td>7,654</td>
<td>9,059</td>
<td>8,548</td>
<td>7,966</td>
<td>8,321</td>
<td>41,548</td>
</tr>
<tr>
<td>The average amount of references</td>
<td>38.27</td>
<td>45.30</td>
<td>42.74</td>
<td>39.83</td>
<td>41.40</td>
<td>41.51</td>
</tr>
</tbody>
</table>
Literature types

Among the cited 41,548 references, the main types of literature are journal articles, conference papers, books, reports, dissertations, standard documents and patent literature, as shown in Figure 7.

Correlation Analysis of Journals

There are more than 400 journals listed in the references. According to the number of the published articles, the first three journals are Mechanical Systems and Signal Processing, Nonlinear Dynamics, and Journal of Sound and Vibration, respectively. According to the cited frequency of papers, nine of the top 50 journals are cataloged into the aerospace discipline, other 41 journals are closely related to the discipline, as shown in Figs. 8 and 9.
Correlation Analysis of Monographs

The statistics analysis on the monographs suggests that the first 50 monographs published in 1959-2009 averagely have a citation of more than 6 times, and most of which have second edition or fourth edition. They are classic books of aerospace discipline. The publication years of these important books are given in Figure 10. One can see from Figure 10 that the most cited books are published in the 1990s, which means that half-life period of books are much larger than that of journals. The top 50 closely related books are shown in Figure 11.
Important Academic Conferences

The conference literature is a very important way to disseminate the latest scientific research achievements. According to the number of the published papers, the top three are *Proceedings of the IEEE, American Control Conference, and IEEE Conference on Decision and Control*, as shown in Figure 12.
Figure 12. Academic meeting TOP30

Important Scientific and Technological Reports

The science and technology report is an important special document of aerospace discipline. With the statistical analysis of the scientific reports cited by the highly cited papers, it is found that the most cited scientific and technical reports are AIAA Paper, SAE Paper and NASA’s Report, as shown in Figure. 13.

Figure 13. Technical report TOP 9

Conclusions

According to highly cited papers and their references, the research hotspots, the international cooperation, and the literatures’ types of a discipline can be found. Based the world-famous citation database Scopus, this paper studied the characteristics of aerospace discipline. The results show the average length of highly cited papers is 13.87 pages, and the average number of references is 41.51; The countries contributed most in highly cited papers
are China, the United States and the United Kingdom; The figure of the high frequency word revealed the academic research hot spots of nearly five years; The paper reveals the highly influential journals, books, science and technology reports and academic conferences related to the subject by the way of charts: The most cited books are published in the 1990s. Half-life period of a book is much larger than that of a journal.

Acknowledgments
I feel grateful to my supervisor, Professor Qin Ping, for her valuable instructions and suggestions on my thesis. I also owe my sincere gratitude to my friends and my fellow classmates who gave me their help.

References

Differences in citation personal display: Does it exist in three social science disciplines?

Xingchen Li\(^1\)  Qiang Wu\(^2\)  Yuanyuan Liu\(^3\)

\(^1\) lxc92@mail.ustc.edu.cn
University of Science and Technology of China, Hefei (China)

\(^2\) qiangwu@ustc.edu.cn
University of Science and Technology of China, Hefei (China)

\(^3\) liuyy604@mail.ustc.edu.cn
University of Science and Technology of China, Hefei (China)

Abstract
Based on researchers’ personal websites, the extents of citation personal display (CPD) in three disciplines of social science (economics, sociology, and psychology) are investigated. Eight universities in the U.S. are targeted according to the 2016 US News & World Report ranking by subject. Results show that the overall level of CPD in social science is not high (7.5%), and researchers in psychology have the highest level of CPD (8.3%), followed by economics and sociology, which means that CPD does not behave the same in three disciplines. Furthermore, the present work suggests that researchers in different universities indeed have differences in CPD when controlling them in the same disciplines, embodying in that the statistically significant differences on CPD exist in economics and psychology, yet differences between universities are insignificant in sociology.

Conference Topic
Citation analysis; Emerging Issues
Introduction

With the development of scientometrics and informetrics, there are a great number of citation-based indices which are presented to evaluate scientists, such as h-index (Hirsch, 2005), g-index (Egghe, 2006), and w-index (Wu, 2010). Previous studies have pointed out that some citation-based indices can be useful in measuring scientists’ achievement (Lehmann et al., 2006; Ball, 2007). Considering that citation itself is a kind of behavior which researchers may want to give recognition to scientists they cited (Merton, 1973; Merton, 1988) or persuade others to support their findings (Knorr-Cetina, 1981; Gilbert, 1977; Collins, 2004), hence, researchers are more likely to show different attitudes towards citations and citation-based indices. Then the intriguing question to ask is whether these indices have been used or adopted among researchers, which remains an open question and brings this study’s value. The primary aim of this research is to investigate the extent to which researchers use citations or citation-based indices in some way.

Since there are many studies shown that the personal website can be a wonderful tool to assess an individual researcher (Ajiferuke & Wolfram, 2004; Máñ-Bleda & Aguillo, 2013; Kousha & Thelwall, 2014), this research wants to study the extent to which researchers use citations or citation-based indices with the help of researchers’ personal websites. Here, the behavior that researchers mention or display citations or citation-based indices or provide a link to Google Scholar Citations (GSC) on the personal websites is defined as citation personal display (CPD) (Li et al., 2017). Li et al. (2017) view citation personal display as a kind of recognition of index validity, and believe that researchers who have CPD may admit the value of citation in highlighting achievements. Hence, studying the degree of citation personal display can help researchers better understand the popularity and values of citations or citation-based indices.

In addition, Li et al. (2017) have studied the extent of CPD with regard to physicists in 11 well-known universities in USA, Britain, and China, and found that the overall proportion of CPD is not high (14.8%). Do researchers in other disciplines also have the low levels of CPD? This will be a question worth studying, which may be able to reflect disciplinary difference on CPD. Based on the above settings, this study tries to answer the question from the perspective of three disciplines of social science (economics, sociology, and psychology). Besides, this paper also intends to analyze the disciplinary differences and tests whether CPD differences are significant across disciplines and universities.

Data and Method

This paper has selected three disciplines of social science (economics, sociology, and psychology) as the objects to investigate the extent to which researchers display citations on their personal websites. Universities in the U.S. are targeted, whose standard of selection is in the top 15 in all three disciplines according to the 2016 US News & World Report ranking by subject. The number of such universities is 8, listing as follows: Harvard University (Harvard), Princeton University (Princeton), Stanford University (Stanford), University of California--Berkeley (UC-Berkeley), University
of California--Los Angeles (UCLA), University of Michigan--Ann Arbor (UM-Ann Arbor), University of Pennsylvania (UPenn), and University of Wisconsin--Madison (UW-Madison).

Then, this study manually examined whether individual researchers provided CPD on personal websites from each university’s economics, sociology, and psychology departments, dividing them by discipline. With regard to the targeted researchers, this study only focused on regular university professors (i.e. full professor, associate professor, and assistant professor). Taken together, the datasets consist of personal websites of 1,004 researchers, among which 367 are economists, 213 are sociologists, and 424 are psychologists.

Results

Difference in disciplines

Table 1 shows the results of CPD in three disciplines. It can be found that, overall, 75 of 1004 researchers provide CPD on their personal websites, accounting for 7.5%, which is far lower than researchers in physics discussed by Li et al. (2017). This reveals that researchers in these three disciplines of social science pay less attention to citations or citation-based indices compared to researchers in science. As Table 1 indicates, researchers in sociology have the lowest level of CPD, only with 5.6% sociologists having CPD. The percentage of CPD in economics is almost the same as in psychology, accounting for 8.2% and 8.3% respectively. As for the significance of disciplinary differences on CPD, the chi-square test suggests that it is not significant among three disciplines (p > .05).

In addition, Table 1 also gives the shares of CPD which are directly provided on researchers’ personal websites. In economics, among the 30 researchers who display CPD, only 5 economists directly show their citation information. As for researchers in sociology, all the number of CPD derive from the link to Google Scholar Citations. In psychology, there are 10 researchers who directly show citations or citation-based indices on personal websites, with the percentage being highest among three disciplines of social science (2.4%). This reflects that most researchers who mention CPD may not have a strong desire to use or display citation-related information.

Table 1. Proportions of CPD in disciplines of social science

<table>
<thead>
<tr>
<th>Disciplines</th>
<th>Pw</th>
<th>N.CPD</th>
<th>N. direct CPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economics</td>
<td>367</td>
<td>30 (8.2%)</td>
<td>5 (1.4%)</td>
</tr>
<tr>
<td>Sociology</td>
<td>213</td>
<td>12 (5.6%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Psychology</td>
<td>424</td>
<td>35 (8.3%)</td>
<td>10 (2.4%)</td>
</tr>
</tbody>
</table>

Note. Pw: the number of researchers who have personal websites; N.CPD: the number of researchers who have citation personal display; N. direct CPD: the number of researchers who directly give citation personal display

Difference in universities

Noting that Table 1 has suggested that the extent of CPD in three disciplines of social
science is not high, then does every university have a low level of CPD? In order to know the degree of CPD in eight universities, this paper also analyzes distribution of CPD by universities (Table 2). In economics, there are big differences on CPD between universities, reflecting in that the highest level of CPD is up to 16.9% (UM-Ann Arbor), yet the proportion of CPD in UCLA is low to 0%. Economists in UC-Berkeley and UPenn possess a relatively high level of CPD, being 13.0% and 11.8% respectively. The remaining four universities have a relatively low degree of CPD: Princeton (7.3%), Stanford (6.8%), UW-Madison (6.1%), and Harvard (1.8%). The CPD differences between universities are tested, and the result shows that, in economics, the significant differences among the 8 universities indeed exist (p < .05).

Concerning sociology, this study finds that there are two universities in which no researchers mentions CPD, meaning that sociologists in UC-Berkeley and UW-Madison basically do not display or adopt citations or citation-based indices. Sociologists in Stanford have the highest degree of CPD, followed by Princeton (10.5%). The extent to which sociologists display CPD in other four universities is not high and has small gap, which may be the reason why there are not significant CPD differences between eight universities (p > .05). As for researchers in psychology, the highest percentage occurs in UPenn (20.6%), and the opposite case is in UW-Madison (2.7%). It can be seen that all psychologists in eight universities have provided CPD, which leads to the highest degree of CPD in psychology (8.3%). The chi-square test also shows that statistically significant differences among universities appear in psychology (p < .05).

Table 2. The proportions of CPD in eight universities

<table>
<thead>
<tr>
<th></th>
<th>Economics</th>
<th></th>
<th>Sociology</th>
<th></th>
<th>Psychology</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pw</td>
<td>N.CPD</td>
<td>Pw</td>
<td>N.CPD</td>
<td>Pw</td>
<td>N.CPD</td>
</tr>
<tr>
<td>Harvard</td>
<td>56</td>
<td>1 (1.8%)</td>
<td>22</td>
<td>1 (4.5%)</td>
<td>28</td>
<td>3 (10.7%)</td>
</tr>
<tr>
<td>Princeton</td>
<td>55</td>
<td>4 (7.3%)</td>
<td>19</td>
<td>2 (10.5%)</td>
<td>33</td>
<td>1 (3%)</td>
</tr>
<tr>
<td>Stanford</td>
<td>44</td>
<td>3 (6.8%)</td>
<td>21</td>
<td>3 (14.3%)</td>
<td>42</td>
<td>2 (4.8%)</td>
</tr>
<tr>
<td>UC-Berkeley</td>
<td>46</td>
<td>6 (13.0%)</td>
<td>29</td>
<td>0 (0%)</td>
<td>39</td>
<td>6 (15.4%)</td>
</tr>
<tr>
<td>UCLA</td>
<td>40</td>
<td>0 (0%)</td>
<td>36</td>
<td>3 (8.3%)</td>
<td>114</td>
<td>6 (5.3%)</td>
</tr>
<tr>
<td>UM-Ann Arbor</td>
<td>59</td>
<td>10 (16.9%)</td>
<td>30</td>
<td>1 (3.3%)</td>
<td>97</td>
<td>9 (9.3%)</td>
</tr>
<tr>
<td>UPenn</td>
<td>34</td>
<td>4 (11.8%)</td>
<td>26</td>
<td>2 (7.7%)</td>
<td>34</td>
<td>7 (20.6%)</td>
</tr>
<tr>
<td>UW-Madison</td>
<td>33</td>
<td>2 (6.1%)</td>
<td>30</td>
<td>0 (0%)</td>
<td>37</td>
<td>1 (2.7%)</td>
</tr>
</tbody>
</table>

Note. Pw: number of researchers who have personal websites; N.CPD: number of researchers who have citation personal display.

Conclusions and discussions

The present study aims to investigate the extent of citation personal display in three disciplines of social science (economics, sociology, and psychology) with the help of researchers’ personal websites. The findings show that the overall level of CPD is low, only 7.5% of 1004 researchers in question display citations or citation-based indices on their personal websites. Since Li et al. (2017) find that the degree of CPD in physics is 14.8%, hence, the low extent of CPD in this paper reveals that researchers in social science are less likely to use citations or citation-based indices compared to
researchers in physics. In addition, although the disciplinary differences discussed in this work are not statistically significant, it also can find that researchers in psychology own the highest level of CPD (8.3%), followed by economics (8.2%) and sociology (5.6%).

In addition, in terms of university, this study also finds that there are some differences among the eight universities, but its significance depends on discipline. In economics and psychology, the statistically significant differences on CPD exist, yet, for sociology, it does not show significant difference between universities. This hints that different universities’ policies and culture may have influences on researchers’ attention to citations or citation-based indices. It is an interesting phenomenon that the proportions of CPD in a university are not the lowest if the university has the highest level of CPD in any discipline. Taking psychologists in UPenn as an example, its extent of CPD is highest in psychology (20.6%), and UPenn also has the medium-high degree of CPD in economics and sociology. This may lead to such a problem whether researchers are affected by their peers in the same universities or not and how this affects others, which may be a valuable research topic.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No. 71273250).

References


Usage Pattern Analysis of Academic Articles from Two Chinese Journals

Chen Bikun1 Zhong Zhouyan2 Zhan Changjing3

1 chenbikun@njjust.edu.cn
Nanjing University of Science and Technology, Nanjing (China)

2 594096489@qq.com
Nanjing University of Science and Technology, Nanjing (China)

3 1308205824@qq.com
Nanjing University of Science and Technology, Nanjing (China)

Abstract
Usage metrics have become increasingly popular in scientometric with the advent of electronic resources. While most current researches relied on the usage data retrieved from English publishers. In this study, the usage data was extended to Chinese publishers. We collected and analysed the usage data of two well-known Chinese OA (Open Access) journals in the field of multidiscipline and library and information science respectively from journals’ official websites and Chinese Academic Journals Full-text Database (CJFD) in China National Knowledge Infrastructure (CNKI). It was found that websites’ information architecture affected usage pattern. “Progressive rule”-based site guided users to view full-text articles more than download them and “parallel rule”-based site guided users to download full-text articles more than view them. It also depicted that usage data of most articles was at their peak around pagination month and at the bottom at very earlier or later months. Articles published online as soon as they were peer reviewed did not get the attention of the community unless they were assigned to an issue. The average usage data in both journals overall started to decline after pagination month and then continued to sustain a steady level for several months.

Conference Topic
Usage data; Science of science

Introduction
With the advent of electronic resources, particularly electronic journals, and their increasing acceptance, this resulted in a rapid change in the user preference especially since 2002 (Kraemer 2006; Schloegl & Gorraiz 2010). Usage metrics have become increasingly popular in scientometric analyses beyond librarian practices (Glänzel & Gorraiz 2015). However, most researches up to now were based on the usage data retrieved from English publishers (e.g. Elsevier, Springer, Nature, IEEE Xplore, PLoS and PeerJ) or English citation index databases (e.g. Web of Science and Scopus). In reality, publishers and citation index databases of other languages also began to provide academic usage data. What’s the usage pattern of non-English academic articles? Whether non-English academic articles share the same pattern with English ones? What’s the unique pattern of non-English academic articles? Answering these questions could extend the vision from English platform to non-English platforms, which would provide a new perspective to promote the research on usage metrics. So, the purpose of this study was to explore the usage pattern of non-English academic articles by analysing the usage data of two well-known Chinese OA journals.

Literature Review
The definition of usage may be rather broad. Citation could be regarded as an official, real and essential form of usage of previous studies, while views and downloads are more of unofficial,
potential, unessential/peripheral forms of usage (Wang et al. 2016). In fact, usage metrics are even older than citation metrics, because librarians have been tracking usage since the beginning of their profession, ranging from basic user surveys to the usage tracking of physical journal issues and monographs to library loan statistics to the sophisticated analysis of e-media usage (e-metrics) (Glänzel & Gorraiz 2015).

Different from the librarian practices, usage metrics in Scientometrics mainly focused on the following topics. Firstly, the obsolescence of articles (Moed 2005; Kurtz et al. 2005a, b; Kurtz & Bollen 2010; Liu et al. 2011; Gorraiz et al. 2014; Wang et al. 2014a, b; Moed & Halevi 2016). Secondly, usage metrics were introduced as an alternative to traditional citation metrics (Bollen et al. 2005, 2008; Brody et al. 2006; Duy & Vaughan 2006; Rowlands & Nicholas 2007; Wan et al. 2010), in fact they should rather be regarded as supplementary metrics jointly with altmetrics measures (Barjak, Li, & Thelwall, 2007; Adie & Roe, 2013; Taylor, 2013). Thirdly, usage metrics were perceived as the indicator to evaluate the performance of journals, authors, groups and countries (Bollen et al. 2005; O’Leary 2008; Wan et al. 2010; De Sordi et al. 2016) and to identify the latest research trends (Bollen et al. 2003; Wang et al. 2013b). Finally, usage metrics were used as the approach to explore user behaviors (Davis & Solla 2003; Davis & Price 2006; Wang et al. 2012, 2013a, b; Khan & Younas 2017).

All the researches above were mostly based on the usage data from English publishers or English citation index databases, ignoring non-English publishers and citation index databases. For this reason, this study expanded the English usage data to Chinese usage data to explore their usage pattern. In this study, we were interested in following research questions in the context of two selected Chinese journals in the field of multidiscipline and library and information science respectively:

1. Whether two selected Chinese journals share the same usage pattern? And whether both journals share the same usage pattern with English journals?
2. What’s the correlation among different usage data in the same journal?
3. What’s the pattern of usage data before and after the issue/pagination date?

Data

Due to limited availability of dynamical data, we applied convenience sampling instead of random sampling for our data collection. In this study, we focused on two well-known Chinese OA journals in the field of multidiscipline and library and information science respectively, i.e., *Chinese Science Bulletin* (CSB) (New Site: http://engine.scichina.com/publisher/scp/journal/CSB) (Legacy Site: http://csb.scichina.com:8080/CN/volumn/home.shtml) and *Data Analysis and Knowledge Discovery* (DAKD) (www.infotech.ac.cn). The selected journals published academic articles on their official websites and the full-text articles were also indexed by CJFD in CNKI. So, the usage data in this study were collected from their official websites and CJFD.

*Chinese Science Bulletin* is a multidisciplinary academic journal supervised by the Chinese Academy of Sciences (CAS) and co-sponsored by the CAS and the National Natural Science Foundation of China (NSFC). CSB is a ten-day international journal publishing high-caliber, peer-reviewed research on a broad range of natural sciences and high-tech fields on the basis of its originality, scientific significance and appeal to a general audience. CSB provided monthly downloads data in recent 12 months for each article starting from April 2016.

*Data Analysis and Knowledge Discovery* (formerly *New Technology of Library and Information Service*), being sponsored by CAS and published monthly by the National Science Library of CAS, is one of the leading academic journals in library and information science and related fields from China. DAKD devotes itself to the study and application of the theories, techniques, practices using big data to support knowledge discovery for decision & policy
making. DAKD started to provide monthly downloads data for each article starting from November 2015.

Founded in June 1996 by Tsinghua University and Tsinghua Tongfang Holding Group, CNKI project has been dedicated to promote large-scale digitization of China knowledge resources and establish a platform for global dissemination and value-added services for nearly two decades. CNKI is mainly engaged in integrating the digital publishing of Chinese academic and educational literature resources, providing specialized knowledge services, and undertaking technology development such as digital publishing, content resources management and value-added information services. With complete coverage of China journals, doctoral dissertations and master theses, newspapers, conference proceedings, yearbooks, reference books, encyclopedias, patents, standards, scientific and technological research findings, laws and regulations, “China Integrated Knowledge Resources Database” is the most comprehensive knowledge system that integrates 90% of China knowledge and information resources (http://oversea.cnki.net/kns55/UserGuide/en/index.html). CNKI is a non-OA platform, authorization is required to use its service.

For CSB, only those issues (from issue 11 to 36) that were published from April 2016 to December 2016 were selected because CSB launched its new official site in April 2016 and its legacy site no longer updated new issues after April 2016. And the usage data of the two sites were different and it was hard to integrate them. We manually collected monthly usage data (full-text and abstract views of HTML formats, full-text downloads of PDF format) of these articles from the journal official website and total downloads information (full-text downloads) along with other metadata (e.g., article titles, authors, total cited frequency, keywords, etc.) from CJFD in CNKI. CSB updated its usage data daily, so all the data were downloaded on March 31, 2017 and revised on April 1, 2017. We found that CSB provided “early access” to their content and published some articles online as soon as they were reviewed. Then they were given pagination and assigned to an issue in later dates. Therefore, for CSB, each article had two dates, i.e., “publication date” (early access date) and “current version date” (pagination date/issue assigned date). After deleting the editorials, interviews, news and invalid records, 299 articles were kept.

For DAKD, only those issues which were published from January 2016 to December 2016 were selected because the usage data before 2016 were incomplete. We manually collected monthly usage data (full-text and abstract views of HTML formats, full-text downloads of PDF format) of these articles from the journal official website and total downloads information (full-text downloads) along with other metadata (e.g., article titles, authors, total cited frequency, keywords, etc.) from CJFD in CNKI. DAKD updated its usage data monthly, so all the data were downloaded on April 1, 2017. After deleting the editorials and news, 139 articles were kept.

Results

Statistics of usage data between CSB and DAKD

In Table 1, it showed that users preferred to download CSB articles on CNKI than its official site and view full-text articles more than downloading them on CSB official site. Conversely, users preferred to download DAKD articles on its official site than CNKI and download full-text articles more than viewing them on DAKD official site.

After visiting the official websites of CSB and DAKD, it was found that the differences of websites’ information architecture maybe the main cause to the opposite user preference. On CSB site, the easiest way to read each article is to click the title. After you click the title of each paper, the default link is the full-text html page. But on DAKD site, you can directly click any link (abstract, full-text in HTML or PDF format) you want at first glance in a page. In other
words, CBS site mainly follows “progressive rule” and DAKD site follows “parallel rule”. They revealed user preference of OA journals in a different way. Wang et al. (2014a) proposed that PDF tended to be the preferred format if researchers wanted to print the article or just save in hard discs for later study. User preference on DAKD, the “parallel rule”-based site, confirmed the viewpoints above.

As to “CNKI Download” data, it proved its popularity in China. Because CNKI is a non-OA platform and has a time delay in indexing full-text articles, but its average downloads were more than CSB and slightly less than DAKD.

<table>
<thead>
<tr>
<th>Journal</th>
<th>Type (total number)</th>
<th>Max</th>
<th>Min</th>
<th>Mean</th>
<th>Median</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSB</td>
<td>CNKI_Download</td>
<td>908</td>
<td>5</td>
<td>111.37</td>
<td>80</td>
<td>106.63</td>
</tr>
<tr>
<td></td>
<td>Site_HTML_View</td>
<td>3124</td>
<td>7</td>
<td>102.44</td>
<td>72</td>
<td>196.17</td>
</tr>
<tr>
<td></td>
<td>Site_PDF_Download</td>
<td>184</td>
<td>1</td>
<td>24.72</td>
<td>19</td>
<td>22.21</td>
</tr>
<tr>
<td></td>
<td>Site_Abstract_View</td>
<td>101</td>
<td>0</td>
<td>10.4</td>
<td>4</td>
<td>16.08</td>
</tr>
<tr>
<td>DAKD</td>
<td>CNKI_Download</td>
<td>310</td>
<td>12</td>
<td>96.63</td>
<td>78.5</td>
<td>63.93</td>
</tr>
<tr>
<td></td>
<td>Site_HTML_View</td>
<td>61</td>
<td>0</td>
<td>7.26</td>
<td>5</td>
<td>9.12</td>
</tr>
<tr>
<td></td>
<td>Site_PDF_Download</td>
<td>718</td>
<td>27</td>
<td>119.00</td>
<td>110</td>
<td>78.15</td>
</tr>
<tr>
<td></td>
<td>Site_Abstract_View</td>
<td>91</td>
<td>4</td>
<td>12.51</td>
<td>9</td>
<td>10.07</td>
</tr>
</tbody>
</table>

Correlation analysis of usage data between CSB and DAKD

In Table 2 and 3, Spearman’s correlation coefficient among different usage data in the same journal were investigated. For CSB and DAKD, there was only significant correlation between full-text HTML view and PDF download data in CSB site. Other usage data showed no correlation with each other. It was found that usually most of the downloads take place in the first 2 years after publication (Schlogl et al. 2014). Maybe the datasets were mainly range from January 2016 to December 2016, the current usage data were not accumulated enough to fully reveal their correlations. As to significant correlation between full-text HTML view and PDF download data from CSB site, it further revealed the effects of “progressive rule”, the more full-text views, the probably more full-text downloads.

<table>
<thead>
<tr>
<th></th>
<th>CNKI_Download</th>
<th>Site_HTML</th>
<th>Site_PDF</th>
<th>Site_Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNKI_Download</td>
<td>1</td>
<td>0.484493</td>
<td>0.3721405</td>
<td>-0.1317974</td>
</tr>
<tr>
<td>Site_HTML</td>
<td></td>
<td>1</td>
<td>0.8503451</td>
<td>0.2487813</td>
</tr>
<tr>
<td>Site_PDF</td>
<td></td>
<td></td>
<td>1</td>
<td>0.4274535</td>
</tr>
<tr>
<td>Site_Abstract</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

All results are significant (p-value < 0.05).

<table>
<thead>
<tr>
<th></th>
<th>CNKI_Download</th>
<th>Site_HTML</th>
<th>Site_PDF</th>
<th>Site_Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNKI_Download</td>
<td>1</td>
<td>0.3599784</td>
<td>0.456298</td>
<td>-0.05239672</td>
</tr>
<tr>
<td>Site_HTML</td>
<td></td>
<td>1</td>
<td>0.3435775</td>
<td>-0.11266621</td>
</tr>
<tr>
<td>Site_PDF</td>
<td></td>
<td></td>
<td>1</td>
<td>0.11784133</td>
</tr>
<tr>
<td>Site_Abstract</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Most results are significant (p-value < 0.05) except the significance test between CNKI_Download and Site_Abstract (p-value = 0.539).
*Distribution of accumulated monthly usage*

The Figure 1, 2 and 3 depicted the accumulated monthly usage distribution (full-text view, pdf download and abstract view) of articles published in CSB and DAKD respectively. Each curve represented one paper and the starting point of each curve represented its published month. The red lines represented most viewed and downloaded articles. For most papers, the growth of accumulated usage showed slowly upward trend after the initial rapid growth period. Only a few papers continued to represent a new jump after the initial rapid growth. Among the most viewed and downloaded articles (denoted by red lines), most of them were published in earlier months than later months. The reason was obvious that articles published in earlier months complete more month accumulations than later months. But Attention should be paid to the most viewed and downloaded articles in later months because after several months they might receive more views and downloads. Besides, Figure 1, 2 and 3 showed the effects of websites’ information architecture on usage pattern between CSB and DAKD in the perspective of time series.

![Figure 1 Distribution of accumulated monthly HTML view between CSB and DAKD](image1)

![Figure 2 Distribution of accumulated monthly PDF download between CSB and DAKD](image2)
Figure 3 Distribution of accumulated monthly Abstract view between CSB and DAKD

Average usage distribution in the month of pagination

To understand the trends both prior and posterior to the pagination month, we aligned the usage data of articles with respect to their respective pagination months. Figure 4 showed the average usage distribution of articles with respect to pagination month for both CSB and DAKD. It depicted that usage data of most articles was at their peak around pagination month and at the bottom at very earlier or later months. Also, it can be observed that in both journals the average usage increases rapidly in the month of pagination except “DAKD PDF download” and “DAKD Abstract view”. After tracing the details of “DAKD PDF download” data, it was found that DAKD usually published its papers on 25th of each month. Obviously, usage data of five or six days in the month of pagination were usually less than the following whole month. As to “DAKD Abstract view” data, maybe users preferred to read newly published full-text articles than abstracts for OA journals.

Moreover, it showed that even articles were published online as soon as they were peer reviewed, but they did not get the attention of the community unless they were assigned to an issue. This finding was in accordance with Wang et al. (2014b) and Khan & Younas (2017), where authors found a rise in average usage data of new “issued” articles as compared to early access articles. The authors suggested that this might be due to the fact that publishers send email alerts to their subscribers which ultimately lead to more attention to newly issued articles. This reason seemed to be true in the case of CSB and DAKD. It is interesting that both Chinese journals also send alerts by journal “WeChat Subscription”, a popular public platform for any person or organization to promote themselves in China. Journal “WeChat Subscription” alerts maybe another way to attract their subscribers to newly issued articles.

Besides, the average usage data in both journals overall started to decline after pagination month and then continued to sustain a steady level for several months. This finding was also in accordance with Khan & Younas (2017), where authors suggested that alerts services might be the cause of growth in downloads in pagination month and its subsequent month. Another reason might be that users consider pagination as a kind of quality attribute that articles are fully reviewed and edited.
Conclusion and discussion

In this study, we analyzed the usage data of two Chinese OA journals in the field of multidiscipline and library and information science respectively. It was proved that the websites' information architecture affected the usage pattern. “Progressive rule”-based site guides users to view full-text articles more than download them and “parallel rule”-based site guides users to download full-text articles more than view them.

Also, for CSB and DAKD, there was only significant correlation between full-text HTML view and PDF download data from CSB site. Other usage data showed no correlation with each other. The results might be caused by limited time span of usage data accumulated from 3 to 15 months. But the significant correlation between full-text HTML view and PDF download data from CSB site also revealed the effects of “progressive rule”, the more full-text views, the
probably more full-text downloads. The effects of websites’ information architecture on usage pattern between CSB and DAKD were further confirmed in the perspective of time series analysis.

Finally, it depicted that usage data of most articles was at their peak around pagination month and at the bottom at very earlier or later months. It also showed that articles published online as soon as they were peer reviewed did not get the attention of the community unless they were assigned to an issue. Besides, the average usage data in both journals overall started to decline after pagination month and then continued to sustain a steady level for several months.

There are some limitations in this study. Only 2 journals in the field of multidiscipline and library and information science respectively were investigated, therefore, the conclusions about the usage pattern might be different in other journals or fields, which needed further experiments. Also, the usage data in the study has a limited accumulation period of 3-15 months. A long-term dataset should be tracked and investigated to support the conclusions.

Last but not least, this study focused on Chinese journals. Compared with previous research on international journals, the study is local and relatively limited. Although these limitations exist, we hope that this first try to explore the usage pattern of Chinese journals would provide a new perspective to promote the research on usage metrics.

Acknowledgments
This paper is supported by Youth Program of National Social Science Fund in China (15CTQ035).

References


Research Progress of Library and Information Science: Exploring Chinese Scientific Articles from 2001 to 2015

Liu Hao

liuhao@mail.las.ac.cn

Scientometrics & Evaluation Research Center (SERC), Chengdu Library and Information Center of Chinese Academy of Sciences, Chengdu, 610041 (China)

University of Chinese Academy of Sciences, Beijing 100040 (China)

Abstract

Academic article can reflect the performance of science research. It is necessary to learn the research progress of library and information science (LIS) from the perspective of English articles. This paper regards the articles written by Chinese scholars published during the period from 2001 to 2015 in Web of Science Core Collection as research object, and explores the research progress of LIS from three aspects, which include an overview of Chinese and world articles published in LIS, the most influential institutions and the main research contents. The performance of Chinese scholars in LIS has been enhanced in terms of both quantity and participation in the world. The threshold of the most influential institutions is gradually increased. Besides, the most influential institutions in LIS in China have shifted from Hong Kong to Chinese Mainland. Information retrieval, digital libraries, information systems, data mining, knowledge management and etc. can be considered as classic research contents in LIS. This paper provides the research progress of library and information science in China from three aspects.

Conference Topic

Mapping and visualization; Social network analysis

Introduction

Talent is an important guarantee for the sustained and healthy development of the national economy and society. In recent years, the Chinese government pays attention to the development of education and the cultivation of talents. In 2015, the State Council of China issued an Overall plan: Promoting the Construction of World-class Universities and Disciplines to improve the level of educational development, enhance the core competitiveness, and lay the foundation for long-term development. Library and information science plays an increasingly important role in the human production and life. The future development of Library and information science is particularly important. Considering the growing importance of library and information science, this study seeks to discover the research progress of library and information science in China from the perspective of English articles to offer reference for leaders and policymakers to make policy for scientific development of library and information science, and help Chinese scholars know the current international research fronts. The study is organized as follows: the first section is research background, the purpose and layout of this article. The second section is the status quo of research, which include the progress of LIS education and research. The third section presents data sources and
methodology. In the fourth section, the results are presented from three aspects, which contain an overview of Chinese and world articles published in LIS, the most influential institutions and the main research contents. Finally, it is discussion and conclusion.

The status quo of research
Many Chinese scholars have done researches on the process of library and information science, and have achieved fruitful results.

Progress of LIS education
Based on the actual materials provided by 24 educational institutions, Zhan (2000) analyzed and summarized the achievement and progress in LIS education, and pointed out the existing problems in Chinese LIS education. Wu and Yu (2015) reviewed the development of LIS educational idea, investigated and analyzed the current state and development trends of LIS education in recent five years. The results indicated that the studies mainly centered on the aspects such as curriculum reform, teaching methods transformation and education objects diversity. Empirical methods were widely used, and research topics were various. Besides, they thought the LIS education in China should emphasize the multi-disciplinary and international collaboration and pay attention to the relationships between people, information and technologies.

Progress of LIS research
Li (2012) took the National Planning Office of Philosophy and Social Science Projects as sample to analyze the hotspots, structures and characters of the research progress with keywords frequency analysis and co-words analysis to reveal the periodical changes in the library and information science research in China and to provide decision-making and reference information for the selection of research thesis, project establishment and educational program planning. Liu et al. (2014) reviewed the research progress of library and information science in China from 2010 to 2013 from four aspects, which included theoretical and librarianship research, information organization and retrieval, metrics and journal evaluation, and historical research of library science. The speculative research had been decreasing gradually, whereas LIS scholars were increasingly focusing on empirical studies on specific research questions resulted from LIS practices or misunderstanding of LIS thoughts. To enhance discipline status and influence of LIS, scholars needed to broaden academic horizons and pay more attention to the universal research questions in the interest of the whole society.

The existing researches in China are mostly based on educational institutions or articles published in Chinese. With the development and progress of Chinese education, institutions and Chinese government pay more attention to the international influence. Under this background, Chinese scholars began to pay attention to English articles, especially high impact articles. Therefore, it is necessary to reveal the research process of library and information science in China from the perspective of English article.

Data and methodology
Data sources
Web of Science Core Collection provides researchers, administrators, faculty, and students with quick, powerful access to the world's leading citation databases. Authoritative, multidisciplinary content covers over 12,000 of the highest impact journals worldwide—including open access journals—and over 160,000 conference proceedings. It'll find current and retrospective coverage in the sciences, social sciences, arts, and humanities, with coverage dating back to 1900 (Thomson, 2017).

The research papers from the field of library and information science are the measurement data in this study. In Web of Science Core Collection, “Advanced Research” was selected, the retrieval items were AD=(Peoples R China OR Taiwan OR Hong Kong OR Macao) AND WC=(INFORMATION SCIENCE AND LIBRARY SCIENCE), document types was article, timespan was 2001-2015, and indexes were SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, CCR-EXPANDED and IC. The date of data collection was March 21, 2017. The resultant dataset contains a total of 3,407 records.

Methodology
The main methods or algorithms used in this study are as followed:
Co-occurrence analysis is an important method in LIS. The word co-occurrences reflect the network of conceptual relations from the viewpoint of scientists and engineers active in the field. And the co-word frequency array is used to construct a co-word map that presents the intellectual content of a field (Tijssen and Van Raan, 1994). Now many researches in LIS use this method to explore the research trend of a field. For example, Xu et al. (2016) used co-occurrence analysis to explore interdisciplinary topics of information science.
Science mapping aims to externalize the big picture of science. Its origin can be easily traced back to the pioneering work of Eugene Garfield on historographics of citation, Belver Griffith and Henry Small on document co-citation analysis, and Howard White on author co-citation analysis (Chen, 2013).
In 2002, the Girvan–Newman algorithm proposed by Michelle Girvan and Mark Newman is a hierarchical method used to detect communities in complex systems (Girvan and Newman, 2002). The Girvan–Newman algorithm detects communities by progressively removing edges from the original network. The connected components of the remaining network are the communities. Instead of trying to construct a measure that tells us which edges are the most central to communities, the Girvan–Newman algorithm focuses on edges that are most likely "between" communities. In this study, the Girvan–Newman algorithm is used to cluster the high-frequency keywords, which lay the foundation to discover and interpret research contents in LIS.

Results
The main results are presented in three sections: an overview of Chinese and world articles published in LIS, the most influential institutions and the main research contents.
The growth of LIS articles

Table 1 presents the annual number of this selective set of articles. In 2001, there were 2,315 articles published in LIS. Among them, only 70 (7.60%) articles were published by Chinese scholars. However, in 2015, there were 4,537 articles published in LIS. Among them, 517 (11.40%) articles were published by Chinese scholars. From the above two sets of data, it can be noted that the performance of Chinese scholars in LIS has been enhanced: not only did the quantity increase but also the country’s relative participation in the world from 7.60% in 2001 to 11.40% in 2015. And, more importantly, in the later years, Chinese original articles display a stronger growth rate (7.39-folds) than that of the world (1.96-folds).

To better understanding Chinese performance, a correlation analysis was developed using time and total number of articles as variables. The Pearson coefficient was applied together with coefficient of determination, (r^2) and t test (t). The results indicate r^2 = 95.5%. In other words, 95.5% of the observed increased can be explained by the variable time. The t test (p value < 0.01) indicates that total articles and year of articles are statistically correlated.

<table>
<thead>
<tr>
<th>Publication year</th>
<th>Papers Chinese</th>
<th>World</th>
<th>Chinese/World (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>517</td>
<td>4,537</td>
<td>11.40%</td>
</tr>
<tr>
<td>2014</td>
<td>455</td>
<td>3,773</td>
<td>12.06%</td>
</tr>
<tr>
<td>2013</td>
<td>430</td>
<td>3,614</td>
<td>11.90%</td>
</tr>
<tr>
<td>2012</td>
<td>368</td>
<td>3,497</td>
<td>10.52%</td>
</tr>
<tr>
<td>2011</td>
<td>308</td>
<td>3,379</td>
<td>9.12%</td>
</tr>
<tr>
<td>2010</td>
<td>243</td>
<td>3,226</td>
<td>7.53%</td>
</tr>
<tr>
<td>2009</td>
<td>242</td>
<td>3,085</td>
<td>7.84%</td>
</tr>
<tr>
<td>2008</td>
<td>183</td>
<td>2,869</td>
<td>6.38%</td>
</tr>
<tr>
<td>2007</td>
<td>127</td>
<td>2,730</td>
<td>4.65%</td>
</tr>
<tr>
<td>2006</td>
<td>137</td>
<td>2,524</td>
<td>5.43%</td>
</tr>
<tr>
<td>2005</td>
<td>101</td>
<td>2,542</td>
<td>3.97%</td>
</tr>
<tr>
<td>2004</td>
<td>75</td>
<td>2,126</td>
<td>3.53%</td>
</tr>
<tr>
<td>2003</td>
<td>83</td>
<td>2,264</td>
<td>3.67%</td>
</tr>
<tr>
<td>2002</td>
<td>68</td>
<td>2,376</td>
<td>2.86%</td>
</tr>
<tr>
<td>2001</td>
<td>70</td>
<td>2,315</td>
<td>3.02%</td>
</tr>
<tr>
<td>Total</td>
<td>3,407</td>
<td>44,857</td>
<td>7.60%</td>
</tr>
</tbody>
</table>

The most influential institutions

In this study, the timespan (2001-2015) is divided into 5 time slices, i.e. 2001-2003, 2004-2006, 2007-2008, 2009-2012 and 2013-2015. The knowledge maps of institution network based on time slices are drawn by CiteSpaceIII software. As can be seen from in Figure 1, more and more institutions are involved in the research of Chinese LIS: in
2001-2003, the number of institutions involved in the research of Chinese LIS is 148; however, in 2013-2015, it is up to 664, increased 4.49-folds compared with the number in 2001-2003. Unfortunately, the density of knowledge map has decreased: in 2001-2003, it is 0.0071; however, in 2013-2015, it dips to 0.0015. Therefore, the collaboration and exchange among institutions involved in the research of Chinese LIS should be strengthened in the future.

The most influential institutions in the Chinese LIS are defined as the rank top 5% in each time slice. Table 2 lists parts of most influential institutions with numbers of articles. It can be learn that similar to the number of institutions, more and more institutions become the most influential institutions in the research of Chinese LIS. More importantly, the threshold of the most influential institutions is gradually increased from 9 in the period of 2001 to 2003 to 14 in the period of 2013 to 2015, which indicates that the research of Chinese LIS is always in progress..

<table>
<thead>
<tr>
<th>Timespan</th>
<th>Most influential institutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001-2003</td>
<td><strong>Chinese Univ Hong Kong</strong> (17), City Univ Hong Kong (17), Natl</td>
</tr>
<tr>
<td></td>
<td>Chiao Tung Univ (13), Natl Sun Yat Sen Univ (11), Natl Taiwan</td>
</tr>
<tr>
<td></td>
<td>Univ (11), Univ Hong Kong (11), Chinese Acad Sci (9), Hong Kong</td>
</tr>
<tr>
<td></td>
<td>Univ Sci &amp; Technol (9) and Tsing Hua Univ (9)</td>
</tr>
<tr>
<td>2004-2006</td>
<td><strong>City Univ Hong Kong</strong> (26), Univ Hong Kong (18), Natl Chiao Tung</td>
</tr>
<tr>
<td></td>
<td>Univ (17), Chinese Univ Hong Kong (15), Natl Taiwan Univ (13),</td>
</tr>
<tr>
<td></td>
<td>Hong Kong Univ Sci &amp; Technol (11), Natl Chengchi Univ (11), Natl</td>
</tr>
<tr>
<td></td>
<td>Sun Yat Sen Univ (11), Chinese Acad Sci (9) and Natl Cent Univ</td>
</tr>
<tr>
<td>2007-2009</td>
<td><strong>City Univ Hong Kong</strong> (40), Hong Kong Polytech Univ (30), Natl</td>
</tr>
</tbody>
</table>

Figure 1 Knowledge maps of institution network based on time slices

The most influential institutions in the Chinese LIS are defined as the rank top 5% in each time slice. Table 2 lists parts of most influential institutions with numbers of articles. It can be learn that similar to the number of institutions, more and more institutions become the most influential institutions in the research of Chinese LIS. More importantly, the threshold of the most influential institutions is gradually increased from 9 in the period of 2001 to 2003 to 14 in the period of 2013 to 2015, which indicates that the research of Chinese LIS is always in progress.
Taiwan Univ (29), Wuhan Univ (25), Natl Cheng Kung Univ (18), Natl Chengchi Univ (18), Univ Hong Kong (18), Fudan Univ (16), other 9 universities, etc

Natl Taiwan Univ (58), Wuhan Univ (54), City Univ Hong Kong (45), Chinese Acad Sci (40), Hong Kong Polytech Univ (33), Zhejiang Univ (27), Natl Cent Univ (24), Fudan Univ (23), Natl Cheng Kung Univ (21), Univ Hong Kong (20), other 13 universities, etc.

Wuhan Univ (113), Chinese Acad Sci (61), City Univ Hong Kong (59), Natl Taiwan Univ (57), Nanjing Univ (41), Dalian Univ Technol (32), Peking Univ (31), Zhejiang Univ (31), Natl Cent Univ (30), Tsinghua Univ (29), other 26 universities, etc.

For further describing the evolution process of most influential institutions, this study only selects and compares the top 9 of institutions in each time slice. The stacked column of institution based on time slices is drawn (Shown in Figure 2).

Compared with Table 2, it is easy to determine the time slices that most influential institutions are involved in based on the color of stacked column. As can be seen from Figure 2, among 19 institutions, there are only 2 institutions (City Univ Hong Kong and Natl Taiwan Univ) ranking in top 9 in each time slice. Dalian Univ Technol and Zhejiang Univ from Chinese Mainland firstly rank in top 9 during the period from 2013 to 2015. The influence of Tsing Hua Univ has declined in the research of Chinese LIS. During the period from 2001 to 2003, in the top 9 most influential institutions, most institutions come from Hong Kong; however, during the period from 2013 to 2015, most institutions come from Chinese Mainland. What’s more, among 4 institutions that publish more than 100 articles during the period from 2001 to 2015, two institutions (Wuhan Univ (192) and Chinese Acad Sci (119)) come from Chinese Mainland, however only one institution (City Univ Hong Kong (187)) does come from Hong Kong. In summary, those reflect that the most influential institutions in LIS in China have shifted from Hong Kong to Chinese Mainland.

Figure 2 Stacked column of institution based on time slices

The main research contents
The co-occurrence network of high-frequency keywords in articles can provide an effective way to identify research contents. According to 5 time slices, this paper generates five knowledge maps. The drawing process is as follows: first, use the matrix analysis function provided by the Thomson Data Analyzer (TDA) to generate the co-occurrence max of high-frequency keywords (about top 5% in each time slice). Second, adopt NetDraw to draw the knowledge map of high-frequency keywords. Third, use the GN algorithm in NetDraw to divide the network into clusters. And finally, the main research contents of each time slice are interpreted and decided.

In five knowledge maps, we can determine which nodes are in the cluster based on either the shape or the color of nodes.

The analysis of LIS in 2001-2003

As can be seen in Figure 3, the co-occurrence network of high-frequency keywords occurring in 2001-2003 can be divided into nine clusters. The main research contents of Chinese LIS during the period of 2001 to 2003 can be summarized as information retrieval, digital libraries, electronic commerce, education, information technology, data mining, information systems, telecommunications policy, information management and etc. In this period, Chinese scholars focusing on the acquisition, management and application of data, information and knowledge.

![Figure 3 Knowledge map of high-frequency keywords in 2001-2003](image)

The analysis of LIS in 2004-2006

As can be seen in Figure 4, the co-occurrence network of high-frequency keywords occurring in 2004-2006 can be divided into five clusters. Therefore, during the period of 2004 to 2006, the main research contents of Chinese LIS can be summarized as knowledge management, digital library, information technology, information retrieval, electronic commerce, ontology, e-government, user satisfaction, decision support and etc. In this period, while paying continuously attention to research of the acquisition,
management and application of data, information and knowledge, Chinese scholars focus on the research of user behaviour.

Figure 4 Knowledge map of high-frequency keywords in 2004-2006

The analysis of LIS in 2007-2009

As can be seen in Figure 5, the co-occurrence network of high-frequency keywords occurring in 2007-2009 can be divided into ten clusters. Therefore, during the period of 2007 to 2009, the main research contents of Chinese LIS can be summarized as digital library, information retrieval, knowledge management, customer satisfaction, information service, innovation, data mining, electronic commerce, consumer behavior, government, virtual community, artificial intelligence and etc. In this period, virtual community and artificial intelligence become new focuses of Chinese scholars.

Figure 5 Knowledge map of high-frequency keywords in 2007-2009
The analysis of LIS in 2010-2012

As can be seen in Figure 6, the co-occurrence network of high-frequency keywords occurring in 2010-2012 can be divided into nine clusters. Therefore, during the period of 2010 to 2012, the main research contents of Chinese LIS can be summarized as knowledge management, social network, digital libraries, model, information services, metric, e-government, patent, customer satisfaction and etc. In this period, Chinese scholars focusing on the research of measurement and evaluation, patents research are emerging.

The analysis of LIS in 2013-2015

As can be seen in Figure 7, the co-occurrence network of high-frequency keywords occurring in 2013-2015 can be divided into eight clusters. The main research contents of Chinese LIS can be summarized as metric, knowledge management, collaboration issue, research trend, big data, social media, satisfaction, digital library, e-commerce, virtual community, information quality, service quality, patent, information retrieval and etc. In this period, Chinese scholars focusing on the research of big data and social media.
Figure 7 Knowledge map of high-frequency keywords in 2013-2015(Threshold=2)

The evolution of research contents in LIS

Compared the research contents in five time slices, it is not difficult to find that the size of high-frequency keyword in a cluster becomes bigger with times. In other words, the results of research contents in the cluster are more and more difficult to distinguish. We also learn that with the continuous development and progress of science and technology, the research contents in LIS are highly integrated. Information retrieval, digital libraries, electronic commerce, information technology, information systems, data mining, knowledge management and etc., which appear both in five time slices, can be considered as classic research contents in Chinese LIS. Besides, Chinese scholars in LIS have new concerned research contents in each time slice. For examples, the research of user satisfaction and behavior shows that Chinese scholars begin to pay attention to the humanity factors and emphasize the core value concept of "humanity-oriented". With the continuous development and progress of science and technology, our society has entered the era of web2.0 and big data. The research contents of virtual community and social media in LIS have become new focuses of Chinese scholars. Of course, big data has become an important research topic. The 4V’s of big data (Gupta, Gupta and Mohania, 2012) – volume, velocity, variety and veracity—have caused people's thinking. Therefore, advanced kind of logical and physical storage structures, advanced kind of heterogeneous data structures, new mathematical theories and new models of big data are needed in the present day (Biswas,
Besides, the useful information and knowledge has become the focus of people's measurement and evaluation. In totally, the new research contents in Chinese LIS closely follows the development of science and technology, reflecting the discipline of LIS are getting opening.

Conclusion and discussion

In this study, the English articles in LIS written by Chinese scholars in Web of Science Core Collection are regarded as research object. From three aspects, this paper studies the research progress of library and information science. An overview of Chinese and world articles published in LIS, finding the performance of Chinese scholars in LIS has been enhanced: not only does the quantity increase, but also the country’s relative participation is strengthened in the world. In the research of the most influential institutions knowledge maps finding that there are more and more institutions participating in the research of library and information science. More importantly, the threshold of the most influential institutions is gradually increased. This further indicates that the research of Chinese LIS is always in a process of development and progress. We also learn that the most influential institutions in LIS in China have shifted from Hong Kong to Chinese Mainland. This reflects the progress of education in LIS in Chinese Mainland. Unfortunately, the density of knowledge map has decreased. Therefore, the collaboration and exchange among institutions involved in the research of Chinese LIS should be strengthened in the future. Despite its limitation (such as the results of research contents in the cluster are more and more difficult to distinguish with times), information retrieval, digital libraries, electronic commerce, information technology, information systems, data mining, knowledge management and etc. can be considered as classic research contents in LIS. This reflects the discipline nature of library and information science -- a continuously developmental and opening discipline. In the future, Chinese scholars should pay attention to both classical research content and new research content, and strengthen the international collaboration to increase the international influence of Chinese library and information science.

References


Thomson, R., (2016-6-30). ‘Web of Science Core Collection’.


Disciplinary Differences in the Achievements of the National Natural Science Award in China

Xu Wang¹ Chunhui Tan²

¹82504336@qq.com
Central China Normal University, Wuhan (China)

²tanadan@aliyun.com
Central China Normal University, Wuhan (China)

Abstract
The purpose of this paper is to explore the differences in distribution of disciplines of the National Natural Science Award (NNSA) in terms of year, age and collaboration, and further discuss the causes of these differences. Using the awards data published from the website of the National Office of Science and Technology Award Office, a scientometrics analysis was conducted to evaluate disciplinary differences. Through social network analysis to demonstrate the situation of collaboration. Results show that there are differences in the growth trend of the disciplines of the awards, applied disciplines grow faster. There is no significant difference between the basic and applied disciplines in terms of researchers’ age. The analysis of collaborative reveal that the collaboration of applied disciplines is more common, the number of partners and institutions have obvious advantages. The geographical distribution of the researchers’ institution also has a great effect on collaboration, the eastern part of China is mainly focused on basic disciplines, applied disciplines in South and North China. This paper intends to figure out the future research trends, and promote the disciplines construction in China's natural science field.

Keywords
National Natural Science Award (NNSA); Reward system; Disciplinary differences; Age analysis; Collaboration analysis; Discipline development

Conference Topic
The application of informetrics on evaluation
Mapping and visualization

1 Introduction
With the rapid development of human society, countries are not only competing in economic, military, and cultural fields. Countries around the world have established various scientific incentives to encourage scientific researchers to create, and promote the development of basic research. China also invested a lot of funds to carry out research in the field of natural sciences (Feng and Pei, 2011). The most influential is the Nobel Prize, which granted to people who make a significant contribution to the world, especially in the field of natural science (Seechurn et al., 2012). In addition, there are Abel Prize in mathematics, Wolf Prize in physics, Turing Award in computer science etc (Socrates, 2015; Misa, 2015).
People's Republic of China State Council issued Regulations on National Awards for science and technology in order to reward citizens and organizations that have made outstanding contributions to the advancement of science and technology in May 1999, and set up five national science and technology awards in 2000. These five awards are the Highest National Science and Technology Award, the National Natural Science Award (NNSA), the State Technological Invention Award, the National Science and Technology Progress Award and the International Science and Technology Cooperation Award. The NNSA is granted to citizens who make a major scientific discovery and clarify the natural phenomena, characteristics and laws in basic research and applied basic research. There are three conditions for this scientific discovery: (1) not yet discovered or elucidated, (2) has great scientific value, (3) has been recognized by natural science communities both at home and abroad. It represents China's important contribution to the study of the world's natural sciences. Therefore, analysis of the NNSA is of great significance to the understanding of the current situation and trends in the field of Natural Science in China.

This paper analyzes the disciplinary differences through disciplines distribution of the awards each year, age distribution of researchers of different disciplines and the distribution of the collaboration in different disciplines. Through these differences, the present situation and development trend of China's research in the field of natural science are described.

2 Related research

At present, the study of Natural Science in China is mainly in the analysis of the National Natural Science Foundation of China, and lack of relevant research on the NNSA. Scientometric research on the award is very helpful to understand the potential information associated with the award. Charlton (2007) using a combined metric of Nobel prizes, Fields medals, Lasker awards and Turing awards to find best nations and institutions for revolutionary science. Some researchers analyze the institutions, collaborations and citation data of Nobel prize winners (Schlagberger et al., 2016; Chan et al., 2015; Harzing, 2013). These studies all analyzed the winners of Nobel Prize, so the analysis of the attributes of the winners is very important. Above all, the influence of age on creativity and productivity of scientific researchers is obvious (Sugimoto et al., 2016). For researchers, with the increase of age, there are many factors affect their research work. In the field of natural science, the results of the researchers are inversely proportional to the age (Bonaccorsi and Daraio, 2003). In addition, collaborative research has become the main way to solve the problem and promote the disciplinary development. The researchers can reflect the collaboration situation of the collaborative project, and collaborative researchers will continue to cooperate in depth research. The analysis of collaborative research can be carried out from three perspectives, researchers, institutions and regions (Zhang et al., 2012). Collaboration institutions are familiar with approaches and methodologies (Cummings and Kiesler, 2008), the probability of these institutions to work together again and produce new results is higher. In recent years, Chinese government sponsored a large amount of funds to promote scientific research institutions to deepen collaboration (Zhai et al., 2014).

Basic research is an important part of scientific research (Henard and McFadyen, 2005). The development of disciplines can not be separated from the basic research. The US government points out that it is necessary to maintain a leading position in all scientific
knowledge, and to strengthen the link between basic research and national goals in 1994 (Abelson, 1994). Human knowledge has been applied and verified, further development to the scientific level and form the knowledge system. The knowledge system continues to develop and evolve, and is divided into disciplines according to some common characteristics. There are significant differences in formation mechanism, research object, research methods, application areas in different disciplines. Some researchers have explored the phenomenon of disciplinary differences in scientific research behavior (Siemens et al., 2014; Weller and Monroe-Gulick, 2014; Thelwall and Pardeep, 2016). Some researchers have explored the phenomenon of disciplinary differences in scientific research results (Kalichman et al., 2015; Pull et al., 2016). Gazni and Didegah (2016) examined the association between author bibliographic coupling strength and citation exchange in different disciplines. Therefore, combine discipline difference with scientometrics has become a hot research topic.

3 Data and methods
3.1 Data collection
The data collected from the website of the National Science and Technology Award Office which announced the NNSA from 2000 to 2016. By the statistics, the NNSA has a total of 563 awards. In 2002, 2003, 2006, 2009, 2013, 2014, 2015 and 2016 have one first award, the others are second awards. Figure 1 shows the overall situation of the number of awards in each year. It can be seen that as time goes by, although the number of awards is different each year, but the number of the situation showed a gradual increase trend.

**Figure 1 The number of awards each year**

These awards belong to 23 disciplines, Table 1 classifies this 23 disciplines according to the disciplinary code in evaluation of science and technology award of China National Standard. In addition, these disciplines can be divided into basic disciplines and applied disciplines for the holistic analysis.

<table>
<thead>
<tr>
<th>Disciplinary code and classification of awards</th>
</tr>
</thead>
<tbody>
<tr>
<td>110 Mathematics</td>
</tr>
<tr>
<td>130 Mechanics</td>
</tr>
<tr>
<td>140 Physics</td>
</tr>
<tr>
<td>150 Chemistry</td>
</tr>
<tr>
<td>160 Astronomy</td>
</tr>
<tr>
<td>170 Earth Science</td>
</tr>
<tr>
<td>180 Biology</td>
</tr>
<tr>
<td>460 Mechanical Engineering</td>
</tr>
<tr>
<td>470 Power and Electrical Engineering</td>
</tr>
<tr>
<td>210 Agricultural Science and Technology</td>
</tr>
</tbody>
</table>
3.2 Analysis Methods

The number of awards and disciplines, the physiological age and occupational age of the researchers, regional and institutional distribution were all analyzed by Microsoft Excel 2010. Collaboration frequency of Research institutions and regional were conducted using BibExcel 1.0.0.0 (Liu and Gui, 2016). Input the data to Microsoft Word 2010 according to BibExcel format requirements, counted the frequency of collaboration and generated the co-occurrence matrix. The co-occurrence network graphics were produced by Pajek64 4.10 (Leydesdorff and Schank, 2008), and combined with the map of China by Coreldraw x7. In this paper, only the primary researcher of the awards was considered for the researcher's age (Butler, 2001), that the primary researcher's contribution was greatest. Women account for only 5% of winners' gender, therefore, this article does not analyze the gender characteristics. Some awards are comprehensive, may belong to different disciplines, but there were only classified according to the disciplinary code in this article. In addition, some researchers belong to different research institutions, but they need to use the resources of different institutions in their research and, therefore, classified this situation as institutional collaboration. In the analysis of regional collaboration, only considered the collaboration between different regions.

4 Results

4.1 The distribution of disciplinary differences in awards

In China, NNSA is the highest prize in the field of basic research, and the award has no quantitative limitation for each disciplines at the time of application, the number of awards in various disciplines represents China's comprehensive plan for disciplines development.
Figure 2 The number of awards for high frequency disciplines

Analysis of the distribution of disciplinary differences in awards can reflect the advantages and development trends of the natural sciences. Figure 2 shows the number of awards for high frequency disciplines. First, Chemistry accounted for the largest proportion, several areas of China's cutting-edge research in the field of chemistry have a high level in the world. The proportion of Biology and Material Science are more than 10%. The proportion of basic disciplines is 54.71%, which shows that the research focus in the field of natural science in China is on these aspects. Second, Mathematics, Physics, Earth Science, Basic Medicine, Electronics and Communications Technology are all more than the average level of 4.35%, indicating that China's research on these disciplines is also more attention. In addition, the awards projects also include disciplines with Chinese characteristics, such as Traditional Chinese Medicine and Pharmacy, Civil Engineering and Construction, Water Conservancy Engineering.

Figure 3 The number of awards in basic disciplines and applied disciplines

Figure 3 shows the annual change trend of number of awards in basic disciplines and applied disciplines. Before 2013, the number of basic disciplines are more than applied disciplines, except for 2006. However, the growth trend of applied disciplines is relatively fast, and gradually narrowing the gap with basic subjects disciplines, in 2013 and 2014, both have the same number. Since then the number of awards of applied disciplines more than basic disciplines. China gradually transforms the ideas of discipline development to promoting transformation of basic theory and accelerate the development of applied disciplines to adapt to the international trend of vigorously developing new disciplines. In general, the number of awards for basic disciplines and applied disciplines has a growing trend. This is closely related to China's investment in the Natural Science Foundation, from China National Natural Science Foundation was set up when the annual investment of 80 million yuan, to the current 24.8 billion, 30 years to enhance the 300 times. At the same time, with the improvement of reward system and rating requirements, in order to ensure the quality and level of the project, the number of awards in recent years has slowed down.

4.2 The age distribution and disciplinary differences of the researchers

4.2.1 Annual distribution of age of researchers

The inherent law of science are very complicated, from quantitative to qualitative leap
requires certain knowledge accumulation. It is particularly for scientific researchers’ work. The researchers won the awards in different time. According to the previous work, this paper collects the physiological age and occupational age of researchers when they win awards, and combines with the disciplines to analysis differences. Physiological age refers to the age from birth to win the award, occupational age refers to the age from researchers started research work or received a Ph. D to win the award. Three researchers have passed away when they were won the award, not included in Statistics. The average age of the researchers of the year is shown in Figure 4.

![Figure 4 Average physiological age and occupational age of the researchers per year](image)

It can be seen that the trend of physiological age and occupational age changes is basically the same, which are older in the early time, and maintain a slow steady decrease after the first three years of rapid reduction. This trend reflects Chinese researchers become younger, especially in recent years. The podium is no longer belongs to those elderly people with gray hair, there are more young faces, and give the vitality to scientific research of China.

4.2.2 The physiological age and the disciplinary difference

The physiological age of 548 researchers can be queried. The youngest researcher is 38 years old, and the oldest researcher is 91. There are 20 researchers win awards at the age from 35 to 40, and 22 after the age of 76. A total of 333 people at the age between 41 to 55 years old, accounting for 60.8% of the total number. Previous research found that 80% of the researchers completed their studies before the age of 50 in the original scientific research (Cole, 1979). Therefore, the age of the winners of NNSA is generally high. It is inseparable from the history of China's development. 28.3% of the researchers were born before the founding of a new country. During that period of time the economic capacity was poor in China, and caused some researchers were forced to make a living and had to postpone the time to carry out research work.

By analyzing the researchers’ physiological age of each disciplines, can reflect the disciplinary differences in age of the researchers. In Figure 5, the researchers’ physiological age of Mechanical Engineering, Civil Engineering and Construction, Pasturage and Veterinary Science, Environmental Science Technology and Resources Science and Technology are relatively high, of Astronomy, Water Conservancy Engineering, Information and Systems Science Related Engineering and Technology, Basic Medicine are relatively low, the remaining disciplines are in the range of 4 years. The average physical age of basic disciplines
is 54.8, and the average physical age of applied disciplines is 53.9, so researchers of applied disciplines are younger at the time of awarding.

![Figure 5 The distribution difference of physiological age in disciplines](image)

**4.2.3 The occupational age and the disciplinary difference**

Previous studies have indicated that the occupational age of the researchers can also influence the output of scientific research (Dietz and Bozeman, 2005). There are 543 researchers can be queried. The researchers’ occupational age are in the range of 7 to 65, and 372 people between 16 to 30 years old, accounting for 68.5%.

![Figure 6 The distribution difference of occupational age in disciplines](image)

Combined with the physiological age of the researchers, the age are mainly distributed in younger intervals. In Figure 6, the researchers’ occupational age of Mechanical Engineering, Civil Engineering and Construction, Environmental Science Technology and Resources Science and Technology, are relatively high, of Water Conservancy Engineering, Astronomy, Basic Medicine, Information and Systems Science Related Engineering and Technology, Chemical Engineering, Computer Science Technology are relatively low, the remaining disciplines are in the range of 4 years. The average occupational age of basic disciplines is 28.8, and the average occupational age of applied disciplines is 27.4, so researchers of applied disciplines are younger at the time of awarding, same as the situation of physical age.

**4.3 Collaboration analysis and disciplinary difference**

With the rapid development of science and technology, the communication of disciplines is becoming more and more frequent (Prakasan et al., 2014). Affected by production, economy, society and other factors, the achievement of scientific research is no longer a single and
independent activity, often requires a number of researchers to cooperate with each other (Leclerc and Gagné, 1994). This analysis can reflect the situation of collaborative research and the distribution of research strength in natural science field of China.

### 4.3.1 Number of collaborators and disciplinary differences

By the statistics, the largest number of collaborators is 10. Awards for 5 collaborators are 390, accounting for 69.3% of all awards. There are 51 awards are completed by one researcher, and most of them is done by researcher from Hongkong. There are only 7 awards for more than 6 collaborators. In the collaboration of researchers of Chinese natural science awards, five collaborators are dominant. After statistics, awards of Mathematics and Astronomy are mainly done by a single researcher. The number of researchers in Mechanics is relatively average. Other disciplines’ researchers tend to work with others, and collaboration among 5 researchers is the most common. According to the regulations of the Ministry of Science and Technology of China, the number of applicants for principal researchers is generally no more than 5. So researchers tend to establish a five-researchers group, rational use of the most resources to obtain the best interests. The proportion of basic disciplines' collaboration is 84.64%, applied disciplines is 95.92%. Therefore, the research in the field of applied disciplines needs multiple researchers to complete.

### 4.3.2 Collaboration between institutions and disciplinary differences

Scientific research institutions provide instruments, facilities and working environments for scientific research. In recent years, China has introduced a number of science and technology policy to advocate scientific research institutions to strengthen scientific and technological innovation. Scientific research collaboration can combine different scientific research institutions, support each other and complete the research topic together, and also improve work efficiency. This paper only analyzes the collaboration between institutions, regardless of collaboration within the institution. There are 165 institutions participated in collaboration, among those institutions, Tsinghua University has the highest number of collaboration, which is 23. Most of institutions with higher collaboration frequency were universities.

Using social network analysis to reveal the present situation of collaboration. In Figure 7, each circle represents a institution. The circle with larger area are Peking University, Tsinghua University, Nanjing University, The Hong Kong University of Science and Technology, and so on. It shows that these institutions are in the central position in collaborative network, and have high impact in the field of natural science research. In addition, there are 17 lines with deeper color, it indicates that the collaboration between these institutions is more frequent.
The number of collaborative institutions in different disciplines indicates great different. Due to the small number of awards, Pasturage and Veterinary Science, Information and Systems Science Related Engineering and Technology are completed by one institution. Water Conservancy Engineering are completed by the collaboration between institutions.

The proportion of collaboration of Information Science and Systems Science, Civil Engineering and Construction, Water Conservancy Engineering, Environmental Science Technology and Resources Science and Technology are higher than 50%, and most of them are completed by two institutions. Further more, The proportion of collaboration of applied disciplines (34.12%) is higher than basic disciplines (28.57%). Therefore, the applied disciplines needs more institutions to participate in.

4.3.3 Collaboration between regions and disciplinary differences
In order to explore the key regions of scientific research and collaboration, this paper classify awards base on the local area of the institution, and reflect hot area of China's natural science research.

In Figure 8, the biggest size of circle is Beijing, that participate in the 96 times collaboration, accounted for 45.07% of the total. In addition to universities, there are many institutions of the Chinese Academy of Sciences to participate, and most of the important institutions set up in Beijing. So Beijing is in the absolute core position. The collaboration frequency of HongKong, Shanghai, Nanjing, Guangzhou, Wuhan and Dalian are also relatively more. The line between Beijing and HongKong is widest, because of 18 times collaboration frequency, accounted for 8.45% of the total. The connection between Beijing
and HongKong is very tight, especially in universities. Professors often teach in both places, and there are many excellent researchers to visit each other (Lam and Minghung, 2001). The majority of the awards which belong to Beijing-HongKong are applied disciplines.

Generally, the collaboration of the NNSA has produced cluster effects. The active regions are mainly distributed in the North China (include Beijing, Dalian, Jinan etc.), East China (include Shanghai, Nanjing, Hangzhou etc.) and South China (include Guangzhou, Hongkong, Shenzhen etc.). Collaboration with regional research institutions is more convenient, easier to communicate and reduce the cost of resources. North China and South China are mainly based on applied disciplines, and East China’s basic disciplines slightly more than applied disciplines.

5 Discussion and conclusions
This paper, by means of correlation analysis tools, used of scientometrics methods to explore the related situation of NNSA in China. From the perspective of disciplines, combined with the age of the researchers and collaboration, intended to figure out the future research trends, and promoted disciplines construction and development in China's natural science field. The following conclusions can be obtained in this study.

The number of awards are increasing, and the increasing trend of applied disciplines is more obvious than basic disciplines. The proportion of traditional disciplines is gradually declining, while the emerging applications increase. This trend is closely related to the development of China. Since China's reform and opening up, a large number of basic disciplinary research have been carried out to keep up with the trend of world development. With gradual strengthening of China's economy, the application of computer technology and communication technology is becoming more and more popular, it is urgent for China to carry out further research on these disciplines. Therefore, the development of these emerging disciplines is rapid. At the same time, the development of these emerging disciplines will also promote the progress of traditional disciplines (Schneider, 1982).

The age of researchers showed inverted U-distribution. Indicating that the output of scientific research is mainly concentrated in the younger stage. There is no significant difference between the basic and applied disciplines in terms of researchers’ age. Only in several special areas have some differences. Because of the special background of the researchers, the time their participated in scientific research work were uncertainty. But overall, the age of the winners of NNSA is generally high.

The disciplinary differences of the collaboration are obvious. In the research of the co-researchers, applied disciplines are more inclined to cooperate by a number of researchers. In the research of the co-institutions, applied disciplines are also more inclined to cooperate by a number of institutions. The research of applied disciplines need to combine theory with practice, so it need more collaboration. So Collaboration between institutions become mainstream. In the research of the co-regions, shows three core regions of North China, East China and Southern China. East China is mainly based on basic disciplines, but North China and South China are applied disciplines. There are also core regions of China's economic development. China has accumulated a good economic capacity since the reform and opening up, has promoted the development of scientific research (Shen et al., 2001). At the same time, the development of the economy cannot be separated from the ability of continuous
innovation of scientific research.

Several limitations of this study should be noted. First, in age analysis section, the date of birth of a small part of researchers can not be queried, the time of researchers start research work is also fuzzy in some data, this may affect the results. Second, in discipline classification, there is no consideration of interdisciplinary situation. In addition, due to the lack of collaboration in some disciplines or some regions, may result in deviation due to insufficient data samples. Therefore, further research need to overcome these limitations, consider the work of this paper, establish a common reward analysis system.

Acknowledgments
The authors thank the conference organizers and reviewers for their beneficial suggestions. This paper is supported by the Fundamental Research Funds for the Central Universities in 2016 (Grant No. CCNU16A02014).

References
Scientometrics, 31(3), 261-292.


Evaluating Effects of World-Class University Program Using Propensity Score Matching: Evidence from China’s Project 211

Ou Yufang
365068580@qq.com
Zhejiang Normal University, Jinhua (China)

Abstract
This study evaluated the effects of China’s Project 211, a key program of China’s world-class university policies, on related variables of funded universities’ research, teaching and service. The effects were measured by examining the frequency of academic paper, monographs, R&D projects, postgraduates, doctoral students, technology transfer contracts and technology transfer income of Project 211 by propensity score matching model. The model results indicated Project 211 had produced a positive effect on funded universities’ variables related to research in generally, greatly increased the quantity of funded universities’ academic paper and scientific research projects, but it was helpless to increase the quantity of funded universities’ academic monographs. Besides, the Project had also caused positive effects on funded universities’ variables related to teaching through increasing the quantity of postgraduates and doctoral candidates; however, it generated negative effects on funded universities’ service variables.

Keywords: world-class university; Project 211; propensity score matching (PSM); program effects

Conference Topic
University policy and institutional rankings

Introduction
Raising the educational and academic status and ranking of universities to that of internationally accepted world-class universities (WCUs) has become the goal of developing as well as developed countries around the globe in recent years. As a result, policymakers in those countries have prioritized building world-class universities that would help their countries obtain a superior position in the global competition (Shin, 2009). Against this backdrop, world-class university programs (WCUPs) were frequently created in some Asian and European countries: Project 211 of China in 1993; Brain Korea 21 program of South Korea in 1999 (Shin, 2009); Center of Excellence program of Japan in 2002 (Yonezawa, 2007); Initiative for Excellence of Germany in 2005 (Hur and Bessey, 2013) and d’excellence initiatives of France in 2010 (Cremonini, Benneworth, Dauncey and Westerheijden, 2013) etc.

Although many of WCUPs are young, having started in the past decade or even more recently, they have impacted participating universities in a significant way (Salmi, 2016). This makes it imperative to assess what effects these WCUPs have produced and draw lessons from recent and ongoing experiences. However, there are at least two challenges, i.e. time and attribution, to measure effects of WCUPs on universities (Salmi, 2016). The time challenge requires a thorough analysis looking at a reasonably large sample of institutions for comparison purposes, either within a given country or across countries, over many years. The attribution challenge requires an in-depth evaluation to judge whether WCUPs rather than other factors caused the positive or negative changes in universities (Salmi, 2016).

Studies on effects of WCUPs are relatively few, if any, because of the above mentioned two challenges (Seong, Popper, Goldman, Evans and Grammich, 2008; Hou, 2012; Hou, Ince and Chiang, 2012; Salmi and Froumin, 2013). As most WCUPs focus on research excellence (Kehm, 2013; Zhu, 2014), thus it is not surprising that a flood of literature has demonstrated that WCUPs can effectively improve targeted universities’ research performance. For instance, Jung Cheol Shin (2009) found the growth of research publications from Korean research universities increased significantly following the implementation of the BK 21 project in 1999. After performing a meta-frontier analysis, Yaisawarng and Ying (2014) found research performance of Project 211 universities is better, on average, than of non-Project 211 universities during the period of 2007-2009. Möller, Schmidt and Hornboutstel (2016) found Excellence Initiative had succeeded in concentrating excellent research and fostering
collaborations between universities and the non-university research sector, but had not caused massive changes on the overall German research system.

Because there is a close link between teaching and research, evaluation studies about WCUPs have been involved in universities’ teaching performance. Hou (2012) found 11 universities funded by WCUP of Taiwan in 2006 had been expected to increase their research and improve teaching quality both. The result was echoed by Schmoch, Fardoun and Mashat (2016), who found that building a WCU and a more focus on research did not imply squeezing out of teaching, because new staffs for research were employed, the quality of teaching was even improved. On the contrary, crowding-out effects and stagnation of university teaching were found in WCUPs (Chang, 2013). Kehm (2012) got a similar research result that the importance of teaching was downgraded after the execution of Excellence Initiative in Germany. Thus, it was no wonder that Jamil Salmi (2016) worried about the risk of harmful effects on teaching and learning quality because of the research emphasis of most WCUPs.

While many scholars thought WCUs should be excellent in their third mission, i.e. society services (Shin, 2012; Montesinos, Carot, Martinez and Mora, 2008), there was lack of research done on the relationship between the achievement of WCUPs and the service delivered by WCUs. WCUs have been encouraged to lose their “sense of territorial identity and with … ties to local and regional public support for … educational, research and civic missions” as they sought global recognition (Christopherson, Gertler and Gray, 2014). Performances of research, teaching and service differentiated WCUs from national-class universities and local-class universities (Shin, 2013). The aim of Project 211 was building high-level national universities by improving their capacity in teaching, research and public service (Liang, Wei and Liang, 2016).

Although previous studies have demonstrated that WCUPs can effectively improve universities’ all kinds of performance, research performance especially, a potential problem in empirical studies on this topic regarding targeted universities is ample selection bias documented in Heckman (2013). More specifically, universities with good prior performance are more inclined to be funded by WCUPs. Consequently, observed improved performance may be driven by other reasons, rather than WCUPs themselves. Besides, as it seemed that there was no comparability between funded universities and non-funded universities, most of those studies only focused on funded universities’ performance rather than compared them with non-funded universities. The reason was that most WCUPs were non-random funded projects, that was to say the performance of universities’ three missions, i.e. teaching, research and service, were better in target universities than in non-target universities before the WCUPs started, and Project 211 in China was one of classic example of such performance. If we still directly compare Project 211 universities with non-Project 211 universities under this non-random condition, the compared results will overestimate the effects caused by Project 211 on the one hand, and compared advantages cannot be judged as caused by the Project or by natural development of universities on the other hand. To solve the sample selection bias problem, propensity score matching (PSM) model is utilized in this study.

In this study, the author measures effects of Project 211 on related variables of funded universities’ research, teaching and service. This study attempts to meet the first challenge by gathering a unique dataset that provides detailed information on related variables of research, teaching and service from a set of 71 Project 211 universities and 171 non-Project 211 universities over 6 years covering three phases of the Project. Besides, this study tries to conquer the second challenge by the application of propensity score matching. The analysis is guided by the research question, i.e. does Project 211 truly improve research, teaching and service in participating universities?
PSM Model Specification

Research Hypothesis

In the present study, Project 211 administrators contracted with funded universities to perform multiple tasks, such as conducting research, teaching classes and serving the public, all of which were valuable to the Project. The research performance-based incentive system was adopted in funded universities to motivate themselves to spend more time and efforts on research, thus enhanced their research productivity, but reduced their opportunistic behaviours, such as spending time on their teaching and service. Besides, previous study proved research and teaching were mutual promotion (Hou, 2012; Schmoch, Fardoun and Mashat, 2016), thus the research hypothesis is proposed as below:

**H1** : Project 211 has produced positive effects on variables related to research and teaching but negative effects on variables related to service of funded universities.

Methodology

As sample selection bias must be controlled just as mentioned above, propensity score matching model developed by Paul R. Rosenbaum and Donald B. Rubin (1983) is chosen in this study. Utilizing observable characteristics (covariates), the model integrates various covariates into one variable (dimensionality reduction), divides samples into treatment group and control group and performs an equilibrium treatment that is near randomized on various confounding factors in non-randomized studies after balancing the distribution of covariates between treatment group and control group to reduce selection bias (Koolwal and Gayatri, 2010). Using this approach, this study can obtain propensity scores (PS), which measure the extent of matching of the treatment group and the control group in multi-dimensions.

In the following, this study will briefly introduce how to calculate PS value, followed by discussions regarding the three matching approaches and the average effect of treatment on the treated (ATT).

PS is defined as “the conditional probability of receiving a treatment given pre-treatment characteristics” by Rosenbaun and Rubin (1983).

\[
\text{PS is defined as} \quad \Pr(D = 1|X) = \frac{e^{\beta X}}{1 + e^{\beta X}}
\]

where \(X\) is the multidimensional vector of independent variables which may affect the propensity of universities to get Project 211 funds, and \(\beta\) is the vector of coefficients. PS is the predicted value of the Logit model.

To estimate PS score, following Dehejia and Wahba (2002) and Becker and Ichino (2002), this study uses the Logit model with the following steps.

This study starts with estimating probabilities using the Logit model,

\[
\Pr(D = 1|X) = \frac{e^{\beta X}}{1 + e^{\beta X}}
\]
Sample and Variables

To prove hypothesis H1, this study chooses 71 Project 211 universities and 171 non-Project 211 universities that were listed in the top 200 in the *Chinese Universities and Discipline Evaluation Consult Report* released by Research Center of Chinese Science and Evaluation (RCCSE) in the past 5 years. Those universities are divided into two groups, i.e. (a) treatment group——71 Project 211 universities, and (b) control group——171 non-Project 211 universities. As Project 211 was launched in 1993 and quitted in 2016, this study should have got 23 years’ data covering each year from 1993 to 2015 as a perfect longitudinal study. However, the actual building of Project 211 consists of three phases, i.e. Phase I (1995-2000), Phase II (2001-2005) and Phase III (2008-2011) (Guo, 2012), in addition, it is indeed unlikely that effects on universities would be significantly with the first few years immediately after the beginning of a WCUP (Salmi, 2013), thus data of 2003, 2004, 2006, 2009, 2011 and 2014 is got to calculate ATTs in accordance with the principles of data availability and random sampling.

In accordance with selection requirements of PSM’s covariates and outcome variables and overall goals and selective criteria of Project 211, this study chooses the following indicators in Table 1 as PSM’s variables to evaluate Project 211’s effects.

<table>
<thead>
<tr>
<th>Table 1 Variables of PSM Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>dependent variable</strong></td>
</tr>
<tr>
<td><strong>covariates</strong></td>
</tr>
<tr>
<td><strong>PSM variables</strong></td>
</tr>
<tr>
<td><strong>outcome variables</strong></td>
</tr>
</tbody>
</table>

**Reasons for covariates selection.** The first step in the application of PSM is establishing Logit or Probit model to get PS, which involves in covariates selection, and there are ongoing controversies in the literature as to which variables should be included in the propensity score model. Donald Bruce Rubin and Thomas Neal (1996) thought PSM should incorporate all variables that were related to outcome variables but need not to consider the relationship between dependent variables and outcome variables, while Susan Perkins et al. (2000) proposed to adopt only confounding factors that were related to dependent variables and outcome variables both. This study approves the latter point of view. Therefore, in accordance with Susan Perkins’s view, i.e. covariates should be related to dependent variables and outcome variables both, and the selection criteria on funded universities by Project 211, i.e. own a relatively stable faculty team with high quality; be equipped with advanced level of teaching and research conditions; have a certain number of doctoral programs, national key disciplines
and instruct high-level professional talents; have rich research funds and remarkable research achievements, and make a great contribution to the country with high efficiency in school management; have a significant academic influence at home and abroad; have clear building goals with their own characteristics and guaranteed building funds (National Education Commission, 1993), this study makes building area, library books, teaching and research instrument value, assets value, full-time teacher, teacher with senior title, teacher with secondary title, Yangtze scholar, academicians of the Chinese Academy of Sciences (CAS) and Chinese Academy of Engineering (CAE), doctoral program, national key discipline, technology income and technology expense as PSM’s covariates. The relationship between selection criteria and covariates see the following figure.

Figure 1 the Relationship of Selection Criteria and Covariates

Reasons for research outcome variables selection. The effects of WCUPs on research are difficult to assess, as the outcome of research has a variety of different dimensions. The primary dimensions of research outcome are the performance in terms of publications, provision of doctoral students, contributions to the scientific infrastructure (editorships, leading positions in scientific associations etc.) and the transfer to society (participation in advisory boards, local councils, patent applications etc.) (Schmoch and Schubert, 2009a), among which the publications prove to be a lead indicator (Schmoch and Schubert, 2009b). As academic paper and monograph play a considerable role in publications (Moed and Allg, 2004), and research grants is also a frequent indicator evaluating a university’s research performance (Liefner, 2003; Tammi, 2009), and thus three indicators, i.e. academic paper, monograph and research project of universities are chosen to reflect universities’ research.

Reasons for teaching outcome variables selection. While the components of teaching in universities consists of teachers, measures, students etc., in a broad sense (Wescott, 1978), this study prefers to a narrow sense of teaching that only consists of students, i.e. all kinds of students, including undergraduates, postgraduates, doctoral students and overseas students etc., for evaluation target. Although Project 211 has produced effects on teachers and measures, for instance, Liu, Xianjun (2010) thinks the goal driven by blindly pursuing “world-class”, improper evaluation and assessment system and unscientific teaching management system have affected the perception and disposition of teachers on the relationship between teaching and research. In addition, the measures of teaching must have been influenced by one goal of Project 211 in China’s Tenth Five Plan, i.e. quicken the informatization pace of higher education (State Development Planning Commission, 1995). However, there is no evaluation indicator that directly measures the changes of teachers’ perception or psychology, and thus the effects of those changes are difficult to be assessed. To compensate for this limitation, this study chooses variables related to teachers and measures, i.e. full-time teacher, teacher with senior title,
teacher with secondary title, fixed assets and teaching and research instrument as PSM’s covariates. Besides, considering the teaching goal of Project 211, i.e. make targeted universities become cultivation base of high-level talents in China (State Development Planning Commission, 1995) and many scholars think the higher level of postgraduates and doctoral students be an enhancement in teaching (Schmoch, Fardoun and Mashat, 2016). Thus, this study sets postgraduates and doctoral students as PSM’s teaching outcome variables.

**Reasons for service outcome variables selection.** At present, a consensus on the definition of universities’ service has not been achieved in academic circles. Service has traditionally been mentioned in almost every institution’s mission statement, but less commonly has it been internally and externally clearly defined or fully incentivized (Holland 1997). Weerts and Sandmann (2008) found that institutions tended to frame service as knowledge transfer to the public. Therefore, technology service indicators were chosen as service outcome variables in this study, i.e. technology transfer contract and technology transfer income.

**Empirical Results and Discussion**

*Propensity to be Funded by Project 211*

The first step to perform propensity score matching analysis is estimating PS. PS summarizes several pre-treatment university characteristics of each subject into a single-index, which makes matching subjects on a n-dimensional vector of characteristics feasible for large n. In this study, the study chooses building area, library books, teaching and research instrument value, assets value, full-time teacher, teacher with senior title, teacher with secondary title, Yangtze scholar, academicians of the Chinese Academy of Sciences and Chinese Academy of Engineering, doctoral program, national key discipline and science and technology income as covariates of PSM model. Descriptive statistics of those covariates see the following table.

**Table 2 Descriptive Statistics Characteristics of Covariates**

<table>
<thead>
<tr>
<th>covariate</th>
<th>mean control group</th>
<th>mean treatment group</th>
</tr>
</thead>
<tbody>
<tr>
<td>full-time teacher</td>
<td>1185</td>
<td>1439</td>
</tr>
<tr>
<td>senior teacher</td>
<td>169</td>
<td>256</td>
</tr>
<tr>
<td>secondary teacher</td>
<td>358</td>
<td>467</td>
</tr>
<tr>
<td>Yangtze scholar</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>academician</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>area</td>
<td>66</td>
<td>197</td>
</tr>
<tr>
<td>book</td>
<td>161</td>
<td>197</td>
</tr>
<tr>
<td>asset</td>
<td>90467</td>
<td>129375</td>
</tr>
<tr>
<td>instrument</td>
<td>19952</td>
<td>78754</td>
</tr>
<tr>
<td>doctoral program</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>key discipline</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>technology income</td>
<td>60918</td>
<td>142578</td>
</tr>
<tr>
<td>technology cost</td>
<td>85488</td>
<td>196742</td>
</tr>
</tbody>
</table>

Note: Considering the lagged effect of policies, 6 years of covariates’ raw data is adopted after Project 211 executed in accordance with the principle of random sampling, i.e. 2003, 2004, 2006, 2009, 2011 and 2014.

In line with previous literature (Lian, Foley and Boylan, 2011), the study estimates the Logit model (3) to get PS. Such a multi-dimensional matching is expected to help the study find the control universities that were capable but not selected to the Project 211. The study estimates model (3) with various specifications, and the results are presented in Table 4. It is shown that a university’s probability funded by Project 211 is significantly positively related to full-time teacher, teacher with secondary title, Yangtze scholar, academician of Chinese Academy of Sciences and Chinese Academy of Engineering, teaching and research instrument, doctoral students, national key discipline and science and technology income, but negatively correlated to building
area and fixed asset. The probability to be built by Project 211 is not significantly related to senior teacher and technology cost.

Table 3 The Estimation Results of Logit Models

<table>
<thead>
<tr>
<th>covariates</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m1</td>
<td>m2</td>
<td>m3</td>
<td>m4</td>
<td>m5</td>
</tr>
<tr>
<td>full-time teacher</td>
<td>1.176</td>
<td>2.059</td>
<td>1.547</td>
<td>1.583</td>
<td>0.600</td>
</tr>
<tr>
<td></td>
<td>(-2.65)</td>
<td>(-6.31)</td>
<td>(-5.14)</td>
<td>(-5.28)</td>
<td>(-2.73)</td>
</tr>
<tr>
<td>senior teacher</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(-0.00)</td>
<td>(0.00)</td>
<td>(-0.01)</td>
<td>(-0.05)</td>
<td>(-0.04)</td>
</tr>
<tr>
<td>secondary teacher</td>
<td>1.153</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-3.07)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yangtze scholar</td>
<td>0.609</td>
<td>0.613</td>
<td>0.599</td>
<td>0.608</td>
<td>0.614</td>
</tr>
<tr>
<td></td>
<td>(-5.15)</td>
<td>(-5.2)</td>
<td>(-5.14)</td>
<td>(-5.2)</td>
<td>(-5.27)</td>
</tr>
<tr>
<td>academician</td>
<td>0.754</td>
<td>0.781</td>
<td>0.792</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-3)</td>
<td>(-3.08)</td>
<td>(-3.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>area</td>
<td>-0.911</td>
<td>-0.898</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-4.33)</td>
<td>(-4.32)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>book</td>
<td>-1.109</td>
<td>-0.931</td>
<td>-1.235</td>
<td>-1.264</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-3.80)</td>
<td>(-3.29)</td>
<td>(-4.53)</td>
<td>(-4.63)</td>
<td></td>
</tr>
<tr>
<td>asset</td>
<td>-0.579</td>
<td>-0.534</td>
<td>-0.651</td>
<td>-0.662</td>
<td>-0.756</td>
</tr>
<tr>
<td></td>
<td>(-9.5)</td>
<td>(-6.9)</td>
<td>(-4.44)</td>
<td>(-4.52)</td>
<td>(-5.28)</td>
</tr>
<tr>
<td>instrument</td>
<td>0.342</td>
<td>0.361</td>
<td>0.343</td>
<td>0.364</td>
<td>0.256</td>
</tr>
<tr>
<td></td>
<td>(-2.26)</td>
<td>(-2.36)</td>
<td>(-2.23)</td>
<td>(-2.36)</td>
<td>(-1.76)</td>
</tr>
<tr>
<td>doctoral program</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>(-7.63)</td>
<td>(-7.78)</td>
<td>(-7.92)</td>
<td>(-8.03)</td>
<td>(-7.87)</td>
</tr>
<tr>
<td>key discipline</td>
<td>0.025</td>
<td>0.021</td>
<td>0.021</td>
<td>0.023</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>(-4.32)</td>
<td>(-4.35)</td>
<td>(-4.24)</td>
<td>(-4.22)</td>
<td>(-4.33)</td>
</tr>
<tr>
<td>technology income</td>
<td>0.344</td>
<td>0.354</td>
<td>0.360</td>
<td>0.389</td>
<td>0.357</td>
</tr>
<tr>
<td></td>
<td>(-3.1)</td>
<td>(-3.2)</td>
<td>(-3.2)</td>
<td>(-3.5)</td>
<td>(-3.27)</td>
</tr>
<tr>
<td>technology cost</td>
<td>-0.129</td>
<td>-0.124</td>
<td>-0.16</td>
<td>-0.179</td>
<td>-0.187</td>
</tr>
<tr>
<td></td>
<td>(-1.24)</td>
<td>(-1.20)</td>
<td>(-1.58)</td>
<td>(-1.78)</td>
<td>(-1.91)</td>
</tr>
<tr>
<td>_cons</td>
<td>-5.730</td>
<td>-6.938</td>
<td>-3.615</td>
<td>-3.899</td>
<td>-0.974</td>
</tr>
<tr>
<td></td>
<td>(-3.52)</td>
<td>(-4.51)</td>
<td>(-2.74)</td>
<td>(-2.96)</td>
<td>(-0.84)</td>
</tr>
</tbody>
</table>

Pseudo-R²          | 0.302  | 0.297  | 0.288  | 0.282  | 0.271  |
AUC                 | 0.864  | 0.862  | 0.859  | 0.857  | 0.852  |
Observations        | 1666   | 1666   | 1666   | 1666   | 1666   |

Note: ① The dependent variable of Logit model is “wcu”, a discrete variable, equal to 1 if a university is funded by Project 211, and 0 otherwise.
② ***, ** and * represents 1%, 5% and 10% significance level, respectively, with t-values in parentheses.
③ The AUC denotes the area under the ROC curve.

The final goal of Logit Model is to estimate the propensity scores, according to which the study could match Project 211 universities with their control pairs. Obviously, the model specification is an important thing to ensure the validity of the matching procedure. Unfortunately, there is no straightforward criterion to properly specify the Logit model in the literature. The study uses two diagnostic proxies namely pseudo-R² which is widely used in Logit analysis, and the area under the ROC curve (AUC). The reason this study uses AUC is that, the dependent variable, i.e. “wcu”, in the Logit model is a discrete variable (0/1), while the propensity scores (which is the predicted values of the Logit model) is a continuous variable, thus the traditional statistics (such as Pearson correlation coefficient) could not be used to analyse their correlation (Hosmer and Lemeshow, 2013).
In this case, the pseudo-R2’s value is in the range of 0.271–0.302. Comparing Pseudo-R2 and AUC in Table 5, we can see that specification (1) is better than others. Stürmer, Joshi, Glynn, Avorn, Rothman and Schneeweiss (2006) find that, when we use Logit model to get PS, an AUC value larger than 0.8 can be regarded as a good indicator that the model is well specified. In model (1), the AUC is 0.864, well above the value suggested by Stürmer et al. (2006). Therefore, this study uses model (1) as the basic specification to calculate PS, and then compares university performances between Project 211 universities and non-Project 211 universities.

Sample Matching Results
Austin, Grootendorst and Anderson (2007) find that the quantity of matched sample will increase 20% if PSM is reasonable, and the quantity would be reduced and estimated precision of ATT would be decreased otherwise. As the larger of matched sample, the more precise of ATT, while PSM depends on PS to match controlled universities with treatment universities. Therefore, the quality and reasonableness of matching could be determined by the distribution of the PS kernel density function before and after matching (see figure 2) and the nature of covariates (see table 4). PS value on this figure is calculated by the nearest neighbour method.

![Figure 2 Kernel Density before and after Matching](image)

The following discussion is based on the nearest neighbour matching approach. Figure 2 shows the kernel density functions of the treatment group and the control group, based on pre- and post-matching of the two groups, respectively. Prior studies used all universities in the control group to compare with the treatment group, and thus their results were biased. In contrast, the study chooses universities from the control group to match those in the treatment group, based on propensity scores. After matching, as shown in Figure 2, the kernel density functions of the two groups are much closer, indicating that the characteristics of variables in the two groups are similar after matching. Table 4 below further proves significant differences no longer exist between two group’s characteristics after matching, and thus PSM is reasonable and matching quality is good in this study.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Unmatched Mean</th>
<th>%reduct</th>
<th>t-test</th>
<th>V(T)/V(C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>lnassistant</td>
<td>U 5.9677</td>
<td>54.7</td>
<td>10.14</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>M 5.9677</td>
<td>7.9</td>
<td>85.5</td>
<td>0.195</td>
</tr>
<tr>
<td>scholar</td>
<td>U 0.73499</td>
<td>52</td>
<td>11.67</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>M 0.73499</td>
<td>-1.9</td>
<td>96.4</td>
<td>-0.18</td>
</tr>
<tr>
<td>academician</td>
<td>U 0.12422</td>
<td>28.3</td>
<td>6.27</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>M 0.12422</td>
<td>8.5</td>
<td>70</td>
<td>1.07</td>
</tr>
<tr>
<td>area</td>
<td>U 4.0908</td>
<td>31.5</td>
<td>5.84</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>M 4.0908</td>
<td>4.4</td>
<td>86.1</td>
<td>0.72</td>
</tr>
<tr>
<td>book</td>
<td>U 5.0667</td>
<td>41.5</td>
<td>7.59</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4 the Nature of Covariates between Two Groups before and after Matching
Previous empirical research indicated PSM is effective if reduced standard error (%bias) is less than 3% or 5% (Caliendo and Kopeinig, 2008). In this study, most standard errors are less than 5% after matching, which indicates the balance of covariates is good and common support assumption of PSM has been verified. Besides, the results of paired T-test after matching show the difference of each covariates between treatment group and control group is not significant and proves common support assumption has been verified. At finally, LR chi2 is 11.86 and P value is 0.221 exceeding 0 in the table at the end, which implies a university no longer can be judged from covariates characteristics as a Project 211 university or not after matching, and proves PSM’s common support assumption has been satisfied again.

**Analysis on Effects of Project 211**

To prove H1 hypothesis, the study chooses academic paper, monograph and academic project as indicators that measure universities’ research, postgraduates and doctoral candidates as indicators measuring universities’ teaching, and technology transfer contract and technology transfer income as indicators measuring universities’ service. ATTs of Project 211 on outcome variables after calculating PS through Logit model and verifying PSM, see Table 5.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sample</th>
<th>Treated</th>
<th>Controls</th>
<th>ATTs</th>
<th>S.E.</th>
<th>T-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>paper</td>
<td>Unmatched</td>
<td>1727.02</td>
<td>1101.37</td>
<td>625.65</td>
<td>71.30</td>
<td>8.77***</td>
</tr>
<tr>
<td></td>
<td>matched</td>
<td>1721.79</td>
<td>1521.45</td>
<td>200.34</td>
<td>115.38</td>
<td>1.74*</td>
</tr>
<tr>
<td>monograph</td>
<td>Unmatched</td>
<td>10.64</td>
<td>7.19</td>
<td>3.44</td>
<td>0.65</td>
<td>5.30***</td>
</tr>
<tr>
<td></td>
<td>matched</td>
<td>10.63</td>
<td>10.67</td>
<td>-0.05</td>
<td>0.84</td>
<td>-0.06</td>
</tr>
<tr>
<td>project</td>
<td>Unmatched</td>
<td>991.74</td>
<td>578.53</td>
<td>413.21</td>
<td>39.00</td>
<td>10.60***</td>
</tr>
<tr>
<td></td>
<td>matched</td>
<td>988.63</td>
<td>880.61</td>
<td>108.02</td>
<td>63.89</td>
<td>1.69*</td>
</tr>
<tr>
<td>postgraduates</td>
<td>Unmatched</td>
<td>3890.92</td>
<td>1669.74</td>
<td>2221.18</td>
<td>94.28</td>
<td>23.56***</td>
</tr>
<tr>
<td></td>
<td>matched</td>
<td>3889.57</td>
<td>2310.69</td>
<td>1578.88</td>
<td>148.87</td>
<td>10.61***</td>
</tr>
<tr>
<td>doctoral</td>
<td>Unmatched</td>
<td>714.91</td>
<td>157.95</td>
<td>556.96</td>
<td>21.39</td>
<td>26.04***</td>
</tr>
<tr>
<td>candidates</td>
<td>matched</td>
<td>714.25</td>
<td>260.51</td>
<td>453.74</td>
<td>33.66</td>
<td>13.48***</td>
</tr>
<tr>
<td>transfer</td>
<td>Unmatched</td>
<td>26.66</td>
<td>17.97</td>
<td>8.70</td>
<td>2.96</td>
<td>2.94***</td>
</tr>
<tr>
<td>contract</td>
<td>matched</td>
<td>26.50</td>
<td>43.07</td>
<td>-16.57</td>
<td>4.98</td>
<td>-3.33***</td>
</tr>
<tr>
<td>income</td>
<td>Unmatched</td>
<td>6084.25</td>
<td>2634.32</td>
<td>3449.93</td>
<td>638.55</td>
<td>5.40***</td>
</tr>
<tr>
<td></td>
<td>matched</td>
<td>5989.25</td>
<td>8429.81</td>
<td>-2440.56</td>
<td>1051.14</td>
<td>-2.32**</td>
</tr>
</tbody>
</table>

Note: ****, ** and * represents 1%, 5% and 10% significance level, respectively.

In terms of outcome variables of universities’ research, the above table showed the number of academic paper in treatment group, 1721.79, is greater than in control group after matching, 1521.45; the comparison of academic project shows academic project of treatment group,
988.63, is also greater than that in control group after matching, 880.61; but the ATT of monograph is -0.05 and insignificant (T-value is -0.06), which indicates we cannot judge Project 211 has produced negative or positive effects on monograph. The reason possibly is that the overall context of monograph publications in China is trapped into a vicious circle, i.e. Chinese researchers, facing obstacles of scientific funds, press selection, monograph’s cognition and publication period, have a low intention to publish monograph, which causes printings of monograph are rare but prices are high, and thus the purchasing power of all kinds of libraries is weak. Therefore, Project 211 has produced positive effects on universities’ variables related to research in general, but the effects caused by the Project on monograph are insignificant.

In terms of outcome variables of universities’ teaching, the above table shows the number of postgraduates in treatment group, 3889.57, is greater than that in control group after matching, 2310.69; the comparison of doctoral candidates shows doctoral candidates of treatment group, 714.25, is greater than that in control group after matching, 260.51. This indicates that Project 211 has produces positive effects on Project 211 and non- Project 211 universities both through increasing the quantity of postgraduates and doctoral candidates. This result is greatly possibly caused by one target of Project 211, i.e. make funded universities become the base of high-level teaching (State Development Planning Commission, 1995).

In terms of outcome variables of universities’ service, ATTs of technology transfer contract and technology transfer income are 8.70 and 3449.93 respectively before matching, while -16.57 and -2440.56 respectively after matching, which indicates that Project 211 has produced negative effects on variables related to universities’ service. The possibly reason is that most Project 211 universities are research universities focusing on basic research, while non-Project 211 universities are local universities with orientations of serving the development of their local economy.

In conclusion, the results of PSM show Project 211 has produced positive effects on universities’ variables related to research and teaching, greatly increased funded universities’ academic papers, research grants, postgraduates and doctoral students. However, Project 211 has produced negative effects on variables related to service, i.e. technology transfer contract and income. In a word, Project 211 has produced positive effects on variables related to research and teaching but negative effects on variables related to service in funded universities, and thus Hypothesis H1 to some extent is proved true.

**Conclusion**

The building of a WCU is often commented with deep skepticism, as it is associated with global competition and the pursuit of a superficial gain of reputation, sometimes even hampering a real strengthening of quality (Schmoch, Fardoun and Mashat, 2016). Project 211 of China shows that the decision to become a WCU can be realized with much cautiousness.

This study reveals hypothesis H1 has been proved true. Project 211 in China has produced positive effects on variables related to research, greatly increased funded universities’ academic papers and research grants, which is somewhat consistent with the finding of Yaisawarng and Ying (2014), but is helpless to enhance the quantity of monograph. Although the Project has also produced positive effects on universities’ variables related to teaching, i.e. the quantity of postgraduates and doctoral students, it has produced negative effects on variables related to service, i.e. technology transfer contract and technology transfer income.

This paper contributes to the existing literature in an important way. In term of the overall higher education, time and attribution challenges in WCUPs evaluation (Salmin, 2016) have been solved in a new way. This paper analyses the effects of Project 211 on Project 211 universities and non-Project 211 universities both by PSM which solves attribution challenge. It contributes to the higher education literature by using PSM approach to control for sample selection bias that prior studies in literatures do not control for. In specific, it uses nearest
neighbour matching to identify proper target universities for each of the observations in the sample. Besides, it uses a more recent and extensive data set, covering the period before, at and after of the Project with 6 years of data, which solves time challenge.

The present study has several limitations. Firstly, there are only 71 Project 211 universities in the sample, 2 military medical universities and 39 Project 985 universities that are also funded by Project 211 are excluded from the sample because of data unavailability of the former and weak matching quality of the latter, which would induce the incomplete sample bias. Secondly, as the sample ranges from 2003, 2004, 2007, 2009, 2011 and 2014, which may bias the estimation to some extent. Therefore, further studies should utilize the other proper method, i.e. difference-in-difference model, to evaluate effects caused by Project 985 on 39 universities funded by Project 211 and Project 985 both to compensate for those limitations in the future.

References
Hou, Y. C. (2012). Impact of excellence programs on Taiwan higher education in terms of quality assurance and academic excellence, examining the conflicting role of Taiwan’s accrediting agencies. Asia Pacific Education Review, 13(1), 77-88.


How visible are the research of different countries on WoS and Twitter? an analysis of global vs. local reach of WoS publications on Twitter

Zohreh Zahedi
1
1z.zahedi.2@cwts.leidenuniv.nl
Centre for Science and Technology Studies, Leiden University, Leiden (The Netherlands)

Rodrigo Costas
2
2rcostas@cwts.leidenuniv.nl;
Centre for Science and Technology Studies, Leiden University, Leiden (The Netherlands);
Centre for Research on Evaluation, Science and Technology (CREST), Stellenbosch University, Stellenbosch (South Africa).

Abstract:
The country of authors of 5.9 million Web of Science (WoS) publications with DOI from the years 2012 to 2015 have been compared with the country of Twitter users tweeting these WoS publications in order to study the main scholarly users of Twitter across 10 different countries. For this purpose, the visibility of country’s publications in the WoS and geographical distribution of Twitter users tweeting WoS publications have been analysed. The aim is to study how do they differ and what are their preference in tweeting their own vs. other country’s publication. The findings show that in general, US and UK with the highest proportion of outputs in the WoS, are among the main users of Twitter as well. Moreover, except for US, users tweet publications affiliated to other country more than those from their own country. Also, similar to WoS, it seems that altmetric providers are not free of international biases in their coverage and collection of metrics. Finally, various possible reasons on why publications from some countries attract more Twitter users than others have been discussed.

Conference Topic
Altmetrics; Twitter users and usage; Geographical distribution; Altmetric.com; Web of Science (WoS)

Introduction:
Sharing and discussing scholarly content on social media platforms and particularly on Twitter has become popular in recent years. With about 20-30% of Scholarly WoS publications mentioned on Twitter (Robinson- García, et al., 2014), this platform is considered to be a prominent source of altmetrics after Mendeley (Haustein et al., 2014). Moreover, mentions of publications on this platform could be interpreted as an ‘early indicator of attention or publicity’ (Haustein et al., 2014). However, tweets have been proven to have a low correlation with citations (Thelwall, et al., 2013; Costas, Zahedi, Wouters, 2015a). This could be an indication that tweets capture a different type of impact in contrast to citations (Bornmann, 2014). Also, its uptake varies across fields (Costas, Zahedi, Wouters, 2015b) and countries (Alperin, 2015).
The study of the different typologies of Twitter users interacting with the publications has been proposed as a way to better understand Twitter-based metrics (Haustein, Bowman, & Costas, 2015). In this paper we follow-up on the idea of studying the interactions between Twitter users and scientific publications. Thus, here we focus on the analysis of the geographic distribution of scholarly Twitter users in contrast with the countries of the authors.
of the scientific publications. Based on the geographic info of Twitter user accounts captured by altmetric.com, the aim of this paper is to study to what extent Twitter users are tweeting papers from their own or different countries. The following main research question is targeted:

- How are countries tweeting WoS publications in relation to their scientific outputs in WoS?

**Data and methodology:**

We have used a dataset of 5,989,022 million Web of Science publications with DOI published between 2012 and 2015. The publications were matched based on their DOIs with the Altmetric.com database (metrics until Jun 2016) and the Twitter accounts (users) mentioning these papers have been extracted. Altmetric.com provided country information of the Twitter accounts. This information has been analysed in order to study how the country of the Twitter users differs from the country of the authors of the tweeted publications. In this study, we focus on 10 different countries such as United States (US), United Kingdom (UK), Canada (CA), China (CN), Iran (IR), Brazil (BR), South Africa (ZA), Spain (ES), The Netherlands (NL) and Australia (AU). The following indicators have been calculated:

**Proportion of twitter users of a country:**

.  .  .

**Proportion of twitted papers of a country:**

.  .  .

**Proportion of tweets of a country:**

30% (n=1,747,021) of analysed publications (with an average of 6.27 tweets per paper) were mentioned at least once on the Twitter platform by the 1,327,643 distinct tweeter users, both with known and unknown country information. Table 1 shows the general overview of the share of WoS publications affiliated to authors and tweeted by users across the 10 selected countries. US, China, and UK are the most dominant countries in terms of producing WoS publications (28%, 15.1%, and 8.5% of 2012-2015 WoS publications respectively). Regarding the country of the Twitter users, 36.58% and 25.57% of all tweeted publications (n=1,746,933) are from the US and UK respectively. Moreover, 20.19%, and 9.35% of all Twitter users tweeting any of the publications under study come from the US and the UK respectively.
Table 1. Distribution of 2012-2015 WoS publications with DOIs across the 10 countries

<table>
<thead>
<tr>
<th>Country</th>
<th>2012-2015 WoS pubs with doi</th>
<th>Total Twitter users (% of distinct twitter users)</th>
<th>Total 2012-2015 WoS Pubs with DOI twitted (% of distinct tweeted papers)</th>
<th>Total tweets (% ttw)</th>
<th>Mean tweet (mtw)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>5,989,022</td>
<td>1,327,643</td>
<td>1,746,993</td>
<td>10,958,605</td>
<td>6.27</td>
</tr>
<tr>
<td>AU</td>
<td>236,333 (3.9%)</td>
<td>27,811 (2.09%)</td>
<td>124,843 (7.15%)</td>
<td>267,076</td>
<td>2.14</td>
</tr>
<tr>
<td>BR</td>
<td>139,483 (2.3%)</td>
<td>9,116 (0.69%)</td>
<td>36,941 (2.11)</td>
<td>53,413</td>
<td>1.45</td>
</tr>
<tr>
<td>CA</td>
<td>269,436 (4.5%)</td>
<td>44,070 (3.32%)</td>
<td>155,398 (8.90%)</td>
<td>342,956</td>
<td>2.21</td>
</tr>
<tr>
<td>CN</td>
<td>905,021 (15.1%)</td>
<td>1,869 (0.14%)</td>
<td>8,180 (0.47%)</td>
<td>9,989</td>
<td>1.22</td>
</tr>
<tr>
<td>IR</td>
<td>93,431 (1.6%)</td>
<td>341 (0.03%)</td>
<td>1128 (0.06%)</td>
<td>1,351</td>
<td>1.20</td>
</tr>
<tr>
<td>ES</td>
<td>225,456 (3.8%)</td>
<td>33,469 (2.52%)</td>
<td>148,945 (8.53%)</td>
<td>356,861</td>
<td>2.40</td>
</tr>
<tr>
<td>NL</td>
<td>162,121 (2.7%)</td>
<td>17,435 (1.31%)</td>
<td>71,584 (4.10%)</td>
<td>132,559</td>
<td>1.85</td>
</tr>
<tr>
<td>UK</td>
<td>509,663 (8.5%)</td>
<td>124,081 (9.35%)</td>
<td>446,645 (25.57%)</td>
<td>1,341,452</td>
<td>3.00</td>
</tr>
<tr>
<td>US</td>
<td>1,676,986 (28.0%)</td>
<td>268,060 (20.19%)</td>
<td>638,998 (36.58%)</td>
<td>2,236,403</td>
<td>3.50</td>
</tr>
<tr>
<td>ZA</td>
<td>44,443 (0.7%)</td>
<td>6,465 (0.49%)</td>
<td>28,741 (1.65%)</td>
<td>41,222</td>
<td>1.43</td>
</tr>
</tbody>
</table>

**% of the pubs per country per total pubs defined in the column above is presented in the brackets.
Results:

Country distribution of WoS author and Twitter users

Figures 1 and 2 show that the most dominant countries in terms of publications are the US, China, UK, and Germany. According to Figure 1, 28% of WoS publications are affiliated to the US, 11% to China, followed by 6% to UK and less than 5% of publications are affiliated to other countries (Figure 1).

Figure 1. Distribution of country of affiliation of authors of the 2012-2015 WoS publications with DOI

According to Figure 2, for 42% of Twitter accounts no country information is available. For those Twitter users with geographic information available, 20% of users come from the US and about 10% from the UK while other countries have less than 3% of the Twitter users.

Figure 2. Distribution of country of distinct Twitter users of 2012 -2015 WoS with DOI tweeted publications captured by Altemtric.com (NA refers to Twitter account for which the country information is not available by altmetric.com)
Country of Twitter vs. country of authors

Figure 3 shows whether the country of the Twitter users is the same or different from the country of the authors of the tweeted papers (share of tweeting own vs. other country’s publications). In general, papers affiliated by authors from the US are the most tweeted papers by the users from all the 10 selected countries. For example, out of the 1,128 WoS publications tweeted from Iran, papers from the US (46%), the UK (17%), Canada (9%) and Germany (9%) are the most tweeted ones by Iranian tweeters (n=341 Iranian tweeters). Iranian tweeters tweet only 4% of papers from their own country. Brazilian users tweet their own papers in the fifth place after those by US (44%), UK (16%), Germany (8%) and Canada (8%), and their own country (BR,7%). Dutch, Spanish, and South African publications usually are the third group of the most tweeted publications from their own nationals. British, Australian, and Chinese tweeters have the second place in tweeting their own publications after US publications. American Twitter users have the first place in tweeting their own papers (47%), followed by papers affiliated to the UK (12%), Germany (7%), Canada (6%), and China (6%).

![Figure 3. WoS author country (vertical line) vs. Twitter user country (horizontal line)](image)

Figure 4 shows the relationship between the outputs of each country with the share of the Twitter activity coming from the same country (i.e., the proportion of tweeted papers by users from a given country). Regarding the output, 28% of the 2012-2015WoS publications with DOIs (n= 5,989,022) are authored by US and in case of Twitter activity, 36.6% of all tweeted WoS publications (n= 1,746,993) are tweeted by users from the US. About 15% of WoS publications have a Chinese author but only 0.47% of all tweeted WoS publications are tweeted by users from China. The third country is the UK having 8.51% of the WoS publications authored by them and 25.57% of all tweeted WoS publications tweeted from the UK. Other countries show both less than 5% presence in the WoS and less than 10% Twitter
activity although the relationship between their output and their activity on Twitter is relatively aligned.

![Figure 4. Difference between 2012-2015 publication’s presence in the WoS and Twitter activity](image)

Conclusions and discussions:
In this study we have compared the country of authors of WoS publications with the country of Twitter users tweeting these publications in order to explore the main scholarly Twitter users across the world. The idea is to study how scholarly tweeters are distributed worldwide, and whether any country biases in tweeting scientific publications similar to those previously observed for citation databases (Van Leeuwen, et al., 2011) exists. A special focus has been paid on how users tweet publications from their own or other countries. The results show that the US and the UK are both high in their proportions of outputs in the WoS and users discussing scientific publications on Twitter. These two countries have been also reported to be the top countries of Twitter users in the study of Haustein and Costas (2015). China is the second most important country in terms of WoS publications, however Chinese tweeters are not very active in discussing scholarly outputs. This may have to do with the low uptake of Twitter among Chinese users as they use their local tool (Weibo). Publications from the US are observed to be the most important set of publications tweeted from non-US tweeters. This can be related to the overall size of the US as a scientific producer. This result is in contrast to that obtained for Mendeley in the study of comparing country of authors and readers of Scopus publications by Thelwall and Maflahi (2015). This study showed that, although the higher proportion of US authors leads to the higher proportion of a given country readers in some fields, it seems that in general, Mendeley users tend to select articles from their own country more often than publications from any other country (Thelwall & Maflahi, 2015).

However, based on the above results, the main question is what explains the difference in the country’s publication reach on Twitter and in the general scholarly Twitter usage across different countries. In other words, it is not clear why publications from some countries attract more Twitter users than others. Possible explanations for this could be due to biases in the coverage of Altmetric.com and WoS towards English sources and international publications (Mas-Bleda & Thelwall, 2016) as well as specific publishers or publications with digital identifiers (DOIs). Hence, it seems that local publications are underrepresented and similar biases towards developed countries as those observed for citation indicators can also be expected for Twitter indicators. Thus, the ideal of altmetrics serving scholars from developing countries (the so-called ‘alternative scholars’) is missed, as argued by Alperin (2013). Other
reasons for the observed differences could include cultural, technological, economical and political differences among countries. Among others, technological infrastructures, levels of user’s access to technology, education, information literacy, information behaviour, etc. could play a role here (Zahedi, 2016). Clearly, factors such as the familiarity with the Twitter platform of citizens from specific countries, as well as the extent to which they are oriented, adopted or motivated to use this platform for sharing and disseminating scholarly outputs could have an influence on a country’s scholarly Twitter usage. Hence, all these factors need further investigation on how they could influence scholarly Twitter usage across countries. In addition, understanding the different reasons and motivations of users from different countries to interact with scientific results, could give a better insight of the true potential and relevance of altmetric indicators to study national differences in the reception of scientific publications on social media platforms.

Acknowledgement:
This work was supported by funding from the DST-NRF Centre of Excellence in Scientometrics and STI Policy (South Africa).

Bibliography


The diffusion of medical knowledge in social media - An empirical investigation based on PLOS ALM data

Yu Qi\textsuperscript{1}, Tian Yue\textsuperscript{2}, He Peifeng\textsuperscript{3}, Duan Zhiguang\textsuperscript{4}, and Li Jingyu\textsuperscript{5}

\textsuperscript{1}yuqi351@gmail.com
Shanxi Medical University, Taiyuan(China)

\textsuperscript{2}tianyuebaiyun@sina.com
Shanxi Medical University, Taiyuan(China)

\textsuperscript{3}hepeifeng2006@126.com
Shanxi Medical University, Taiyuan(China)

\textsuperscript{4}dzg52827@aliyun.com
Shanxi Medical University, Taiyuan(China)

\textsuperscript{5}lijingyu67677@163.com
Shanxi Medical University, Taiyuan(China)

Abstract
The study of knowledge diffusion pattern in medical field is of importance to advance medical research, and eventually promote human health. Altmetrics might offer new ways to measure the impact of the sciences beyond the science. By using PLOS ALM data, we investigated the diffusion pattern of medical knowledge in social media. Our results shows that non-medical papers received more attentions than medical ones in terms of Twitter, Mendeley and Counter, however the pageviews and downloads for medical papers on PubMed Central were greater than non-medical ones in term of PMC. ALM scores such as Twitter, Mendeley and Counter shows a sharp increase over time at the very beginning after the papers are published. Then, the increase gradually slows down. China got a relatively low degree of public interest compared to the United States and German. Papers with more content and references tended to have a higher degree of public interest. These findings are informative for the scientists in medical field as well as policy makers and administrators for managing and financing medical research in the future.

Conference Topic
Altmetrics
**Introduction**

Medicine is a discipline closely related to human health. Medical research involves the basic research, applied research, or translational research conducted to aid and support the development of knowledge in the field of medicine. The study of knowledge diffusion pattern in this field is of importance to advance medical research, and eventually promote human health. With the development of social media, the current (medical) academic exchange environment has changed dramatically. As a result, a new term “Altmetrics” was born. Altmetrics was first proposed in September 2010 by Priem, Taraborelli, Groth, & Neylon (2010). Since then, it has received extensive attentions. Several definitions of Altmetrics have been proposed. The common understanding across all definitions is that Altmetrics is a new or alternative metrics to the established metrics for measuring scholarly impact (Erdt, Nagarajan, Sin, & Theng, 2016). Bibliometrics (citation analysis) has always been serving as the gold standard in measuring impact of the sciences. A great drawback of citation analysis is that it only measures the influence of scientific work on science itself. However, Altmetrics (alternative metrics) might offer these new ways to measure the impact of the sciences beyond the science, as it is based on social media platforms and tools, scholarly activities or various user activities in social media environments (Priem, 2014). At this time, Altmetrics may be best conceptualized as complementary and improvement the limitations of traditional metrics rather than replacement (Chisolm, 2016; Costas, Zahedi, & Wouters, 2015; Trueger et al., 2015).

In this paper, we will investigate the diffusion pattern of medical knowledge in social media based on Article-Level Metrics (ALM) data provided by Public Library of Science (PLOS).

**Literature review**

With the booming of social media in the last decades, sharing scholarly materials on social media is becoming increasingly popular (Wang, Fang, & Guo, 2016). Due to the advantages that social media have shown, an efficient channel for scientific communication and dissemination is established, and further strengthened by open access (Wang, Liu, Mao, & Fang, 2015). The spread of research results in social media not only enhances the interaction between scientific community and the public, but also provides a wealth of data for gaining new insights into the research activities. Therefore, many studies grow up. Li, Thelwall, & Giustini (2011) investigates scholarly influence by using CiteUlike and Mendeley data, suggesting that this type of influence seems to be still too small to challenge traditional citation indexes. In addition, Bornmann & Haunschild (2015) use Mendeley data to analyse who reads research articles and which people use which scientific papers. Haunschild & Bornmann (2016) investigates societal impact in scientometrics by using a relatively newly source of Altmetrics data.

With the development of the social media, scholars have more ways to disseminate research than ever before. These venues include open access archives, online journals, and social media tools, such as wikis, blogs, social bookmarking, Facebook, LinkedIn, Twitter and Wikipedia. Wang et al. (2016) found that most of the visits from social
media are generated within only one week after publication. However, social buzz about scholarly articles doesn't last long, which leads to the resulted article visits a rapid decay. In short, easy come, easy go. Allen, Stanton, Di Pietro, & Moseley (2013) quantified the impact of social media release on views and downloads of articles in the clinical pain sciences. They found that the mean±SD rate of HTML views in the week after the social media release was 18±18 per day, whereas the rate during the other three weeks was no more than 6±3 per day. The mean+SD rate of PDF downloads in the week after the social media release was 4±4 per day, whereas the rate during the other three weeks was less than 1±1 per day (p<0.05 for all comparisons).

Medicine, a discipline closely related to human health. However, few attempts were made to study how medical knowledge was diffused in the environment of social media. This study will discover the manifold ways in which medical research is disseminated by using with PLOS ALM data. Specifically, this study will address the following two questions:

- How medical knowledge is disseminated on social media over time?
- What are the differences in topics for papers with different diffusion pattern?

Data and Methodology

Data

PLOS ALM data were used as data source in this study. PLOS ALM, launched in 2009, tracks not only the influence of individual PLOS articles, from times downloaded to mentions in social media and blogs, but also the internal article metrics, including comments, html, readers, like and pdf (Chamberlain, 2013; Lin & Fenner, 2013b; Roemer & Borchardt, 2012). PLOS ALM tracks more than 20 different types of artifacts and collects 5 major categories of impact metrics (Bornmann, 2014; Lin & Fenner, 2013a; Melero, 2015). The ALM API and dataset are freely available for all PLOS articles. Our data in this study were harvested through an open-source application called Lagotto, which retrieves metrics from a wide set of data sources, like Twitter, Mendeley, and CrossRef (Erdt et al., 2016; Kraker, Lex, Gorraiz, Gumpenberger, & Peters, 2015).

We checked the PLOS website for the newly published articles every day and harvest the daily visiting data of each article. We have been collecting the DOI number for all the papers published on PLOS website in December, 2016. The ALM data for each of these papers have been collected through DOI numbers every day from 21 o’clock to 23 o’clock Beijing time since they are published. In total, 2083 papers were collected. The visiting data for each of these papers were recorded and updated daily for 100 consecutive days ever since they were available on PLOS website. Since the PLOS ALM data for articles published in December 9, 13 and 14 were not available for a period of time, they are excluded from this study. Therefore 1588 papers were left for our study. According to PLOS classification criteria, the 1588 papers are grouped into 2 categories, medicine (1205 papers) and non-medicine (383 papers).

The bibliographic data for all these papers were collected from Web of Science, including titles, abstracts, addresses, references and number of pages, etc.
Methods

Article level metrics

Four kinds of ALM scores are used in this study (Table 1). They are Twitter, Mendeley, PMC, and Counter.

<table>
<thead>
<tr>
<th>PLOS ALM</th>
<th>Meaning</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>Number of tweets which share an article on Twitter, a social networking and microblogging service</td>
<td>Social Shares</td>
</tr>
<tr>
<td>Mendeley</td>
<td>Aggregate number of bookmarks of the article added to a Mendeley reader's and group's library.</td>
<td>Academic bookmarks</td>
</tr>
<tr>
<td>PMC</td>
<td>Aggregate pageviews and downloads for the PubMed Central hosted version of the article.</td>
<td>Usage Stats</td>
</tr>
<tr>
<td>Counter</td>
<td>Aggregate pageviews and downloads for an article based on COUNTER 3-compliant collection standards.</td>
<td>-</td>
</tr>
</tbody>
</table>

Characteristic Scores and Scales

A similar method to Characteristic scores and Scales (Glänzel, W., 1988), was used to classify papers into different categories according to Counter scores. The classification procedures are as follows:

- **Step 1**: the mean Counter scores for all the medicine papers are calculated (mean_1).
- **Step 2**: papers with Counter score below mean_1 are classified as “poorly followed” papers.
- **Step 3**: The mean Counter score are calculated for papers with Counter score above mean_1 (mean_2).
- **Step 4**: papers with Counter score below mean_2 are classified as “fairly followed” papers.
- **Step 5**: papers with Counter score above mean_2 are classified as “remarkably followed” papers.

Results

The trend of ALM scores over time

We investigated the changes of average ALM scores for medical and non-medical papers for 100 consecutive days after they are published. As shown in Figure 1, different from the findings by (Erdt et al., 2016; Haustein et al., 2014; Li & Thelwall, 2012; Priem, Piwowar, & Hemminger, 2012), non-medical papers received more attentions from twitter and mendeley than medical ones. However, the page views and downloads for medical papers on PubMed Central were greater than non-medical ones. The results indicate that PMC, hosting journal literature in the field of
biomedical and life sciences, is still the main source for medical scientists to keep updated on newly published medical research findings.

![Graph depicting the trend of PLOS ALM score over time.](image1)

Table 2. The frequency that Counter's increase reaches the peak at nth day (n=1, 2, ..., 10).

<table>
<thead>
<tr>
<th>The nth day after publication</th>
<th>The frequency that Counter's increase reaches the peak at nth day</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 or 2</td>
<td>180</td>
</tr>
<tr>
<td>3</td>
<td>503</td>
</tr>
<tr>
<td>4</td>
<td>178</td>
</tr>
<tr>
<td>5</td>
<td>55</td>
</tr>
<tr>
<td>6</td>
<td>89</td>
</tr>
<tr>
<td>7</td>
<td>90</td>
</tr>
<tr>
<td>8</td>
<td>36</td>
</tr>
<tr>
<td>9</td>
<td>49</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
</tr>
</tbody>
</table>

It can also be seen that ALM scores such as Twitter, Mendeley and Counter showed a sharp increase at the very beginning after the papers were published. Then, the increase gradually slowed down. We calculated the every day’s changes of Counter score for each of the 1205 medicine papers. Table 2 lists the frequency that Counter's increase reaches the peak at nth day. Because the ALM data became available 2 days after the papers was published, the first updates for the ALM data should be the sum of the increase for the first two days. We divided this score by 2, and consider the average score as the Counter increases for the first or second day. For example, the paper titled “Do Elite and Amateur Soccer Players Outperform Non-Athletes on
Neurocognitive Functioning? A Study Among 8-12 Year Old Children” was published on December 1, 2016. The ALM score for this paper was firstly updated on December 3, 2016. So we divide the ALM score (Counter: 38) by 2, then we consider that the Counter score increased by 19 for each of the first two days after this paper was published. As shown in table 2, 503 papers’ Counter increase reached the peak 3 days after they were published, accounting for 41.74% of the total. 90.87% papers’ Counter increase reached the peak within a week. This result indicates that these papers received most intensive attention on social media immediately after they were published.

The topic analysis for papers with different diffusion pattern

By using Characteristic Scores and Scales, all the medicine papers in this study are grouped into 3 categories according to the 3rd day’s ALM data: “remarkably followed” papers(75 papers, Counter>476), “fairly followed” papers (374 papers, 219<Counter<=476), and “poorly followed” papers (756 papers, Counter<=219).

Table 3. The top word frequency of the abstracts for the “remarkably followed”, “fairly followed” papers and “poorly followed” papers.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Words</th>
<th>Frequency</th>
<th>Words</th>
<th>Frequency</th>
<th>Words</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>mutation</td>
<td>57</td>
<td>expression</td>
<td>158</td>
<td>expression</td>
<td>295</td>
</tr>
<tr>
<td>2</td>
<td>cancer</td>
<td>44</td>
<td>model</td>
<td>139</td>
<td>model</td>
<td>266</td>
</tr>
<tr>
<td>3</td>
<td>tumor</td>
<td>25</td>
<td>protein</td>
<td>102</td>
<td>protein</td>
<td>238</td>
</tr>
<tr>
<td>4</td>
<td>resistance</td>
<td>24</td>
<td>mice</td>
<td>101</td>
<td>cancer</td>
<td>157</td>
</tr>
<tr>
<td>5</td>
<td>host</td>
<td>22</td>
<td>blood</td>
<td>83</td>
<td>women</td>
<td>148</td>
</tr>
<tr>
<td>6</td>
<td>expression</td>
<td>21</td>
<td>cancer</td>
<td>72</td>
<td>mice</td>
<td>132</td>
</tr>
<tr>
<td>7</td>
<td>DNA</td>
<td>20</td>
<td>immune</td>
<td>61</td>
<td>blood</td>
<td>128</td>
</tr>
<tr>
<td>8</td>
<td>breast</td>
<td>20</td>
<td>DNA</td>
<td>57</td>
<td>tumor</td>
<td>118</td>
</tr>
<tr>
<td>9</td>
<td>sequencing</td>
<td>17</td>
<td>HIV</td>
<td>54</td>
<td>mortality</td>
<td>110</td>
</tr>
<tr>
<td>10</td>
<td>vaccine</td>
<td>15</td>
<td>host</td>
<td>49</td>
<td>serum</td>
<td>101</td>
</tr>
<tr>
<td>11</td>
<td>steroid</td>
<td>15</td>
<td>children</td>
<td>48</td>
<td>liver</td>
<td>99</td>
</tr>
<tr>
<td>12</td>
<td>ctDNA</td>
<td>15</td>
<td>women</td>
<td>46</td>
<td>receptor</td>
<td>88</td>
</tr>
<tr>
<td>13</td>
<td>sepsis</td>
<td>14</td>
<td>transmission</td>
<td>46</td>
<td>children</td>
<td>85</td>
</tr>
<tr>
<td>14</td>
<td>lung</td>
<td>13</td>
<td>network</td>
<td>45</td>
<td>drug</td>
<td>85</td>
</tr>
<tr>
<td>15</td>
<td>somatic</td>
<td>12</td>
<td>receptor</td>
<td>44</td>
<td>population</td>
<td>83</td>
</tr>
<tr>
<td>16</td>
<td>protein</td>
<td>12</td>
<td>lung</td>
<td>42</td>
<td>resistance</td>
<td>80</td>
</tr>
<tr>
<td>17</td>
<td>pathway</td>
<td>12</td>
<td>ZIKV</td>
<td>42</td>
<td>brain</td>
<td>75</td>
</tr>
<tr>
<td>18</td>
<td>metastatic</td>
<td>12</td>
<td>brain</td>
<td>39</td>
<td>breast</td>
<td>72</td>
</tr>
<tr>
<td>19</td>
<td>genome</td>
<td>12</td>
<td>tumor</td>
<td>39</td>
<td>HIV</td>
<td>72</td>
</tr>
<tr>
<td>20</td>
<td>vaccination</td>
<td>11</td>
<td>mRNA</td>
<td>34</td>
<td>diabetes</td>
<td>67</td>
</tr>
</tbody>
</table>

Abstract analysis
We tokenized the abstracts of all the papers with NLTK, a leading platform for building Python programs to work with human language data. Stop words were removed from the generated tokens. Common words used in medical research were also removed, such as “Disease”, “Patient”, “Cell”, etc. The top word frequencies of the abstracts for the 3 categories are listed in Table 3. “Mutation”, “Cancer”, “Tumor” and “ctDNA” appears most frequently in the abstracts of the “remarkably followed” papers. “Model”, “Mice”, “Blood” are at the top of the word frequency list for both the “fairly followed” and “poorly followed” papers. This results means that cancer research is still the main focus of the medical scientists’ attention, as the problems of curing cancer have not yet been solved.

Author analysis

Table 4. The number of authors from each country for “remarkably followed”, “fairly followed” and “poorly followed” papers.

<table>
<thead>
<tr>
<th>Rank</th>
<th>“remarkably followed” papers</th>
<th>Country</th>
<th>Number of authors</th>
<th>Country</th>
<th>Frequency</th>
<th>“fairly followed” papers</th>
<th>Country</th>
<th>Frequency</th>
<th>“poorly followed” papers</th>
<th>Country</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>USA</td>
<td>USA</td>
<td>112</td>
<td>USA</td>
<td>415</td>
<td>USA</td>
<td>USA</td>
<td>596</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>France</td>
<td>Germany</td>
<td>37</td>
<td>Germany</td>
<td>116</td>
<td>Peoples R China</td>
<td>China</td>
<td>323</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Germany</td>
<td>Peoples R China</td>
<td>22</td>
<td>Japan</td>
<td>104</td>
<td>Germany</td>
<td>178</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Australia</td>
<td>France</td>
<td>21</td>
<td>Australia</td>
<td>69</td>
<td>Japan</td>
<td>162</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>England</td>
<td>Australia</td>
<td>20</td>
<td>Canada</td>
<td>63</td>
<td>France</td>
<td>140</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Japan</td>
<td>England</td>
<td>19</td>
<td>England</td>
<td>63</td>
<td>France</td>
<td>140</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Netherlands</td>
<td>Japan</td>
<td>14</td>
<td>Japan</td>
<td>60</td>
<td>Taiwan</td>
<td>128</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Canada</td>
<td>Brazil</td>
<td>12</td>
<td>Brazil</td>
<td>58</td>
<td>England</td>
<td>126</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>South Korea</td>
<td>Canada</td>
<td>9</td>
<td>Canada</td>
<td>50</td>
<td>South Korea</td>
<td>118</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Italy</td>
<td>Spain</td>
<td>7</td>
<td>Spain</td>
<td>49</td>
<td>Spain</td>
<td>114</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Taiwan</td>
<td>Taiwan</td>
<td>7</td>
<td>Netherlands</td>
<td>43</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Wales</td>
<td>South</td>
<td>6</td>
<td>Korea</td>
<td>41</td>
<td>Brazil</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Belgium</td>
<td>Netherlands</td>
<td>6</td>
<td>Australia</td>
<td>39</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>South Africa</td>
<td>Italy</td>
<td>5</td>
<td>Italy</td>
<td>36</td>
<td>Italy</td>
<td>79</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Spain</td>
<td>Denmark</td>
<td>5</td>
<td>Denmark</td>
<td>26</td>
<td>Iran</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Thailand</td>
<td>Sweden</td>
<td>4</td>
<td>Sweden</td>
<td>24</td>
<td>Belgium</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Austria</td>
<td>Switzerland</td>
<td>4</td>
<td>Belgium</td>
<td>22</td>
<td>Sweden</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Peoples R China</td>
<td>South</td>
<td>3</td>
<td>Africa</td>
<td>21</td>
<td>India</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Finland</td>
<td>Scotland</td>
<td>3</td>
<td>Switzerland</td>
<td>21</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Portugal</td>
<td>Mexico</td>
<td>3</td>
<td>Denmark</td>
<td>18</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
As shown in Table 4, the number of authors from USA ranked No. 1 in all the 3 categories. The number of authors from Germany ranked 3rd, 2nd, 3rd in “remarkably followed”, “fairly followed” and “poorly followed” papers respectively. China ranked 18th in “remarkably followed” papers, while ranked 3rd in “fairly followed” and 2nd “poorly followed” papers.

Country/Region analysis

Table 5. Rank of country’s/region’s productivity for “remarkably followed”, “fairly followed” and “poorly followed” papers.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Frequency</th>
<th>Country</th>
<th>Frequency</th>
<th>Country</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>USA</td>
<td>35</td>
<td>USA</td>
<td>129</td>
<td>USA</td>
<td>216</td>
</tr>
<tr>
<td>2</td>
<td>Germany</td>
<td>9</td>
<td>Peoples R China</td>
<td>41</td>
<td>Peoples R China</td>
<td>118</td>
</tr>
<tr>
<td>3</td>
<td>England</td>
<td>8</td>
<td>Germany</td>
<td>41</td>
<td>Germany</td>
<td>65</td>
</tr>
<tr>
<td>4</td>
<td>Netherlands</td>
<td>5</td>
<td>England</td>
<td>33</td>
<td>England</td>
<td>61</td>
</tr>
<tr>
<td>5</td>
<td>France</td>
<td>5</td>
<td>Netherlands</td>
<td>21</td>
<td>Japan</td>
<td>51</td>
</tr>
<tr>
<td>6</td>
<td>Australia</td>
<td>5</td>
<td>Canada</td>
<td>21</td>
<td>France</td>
<td>44</td>
</tr>
<tr>
<td>7</td>
<td>Canada</td>
<td>5</td>
<td>France</td>
<td>21</td>
<td>Canada</td>
<td>37</td>
</tr>
<tr>
<td>8</td>
<td>South Korea</td>
<td>5</td>
<td>Brazil</td>
<td>19</td>
<td>South Korea</td>
<td>37</td>
</tr>
<tr>
<td>9</td>
<td>Japan</td>
<td>4</td>
<td>Spain</td>
<td>19</td>
<td>Australia</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>Belgium</td>
<td>4</td>
<td>Japan</td>
<td>18</td>
<td>Spain</td>
<td>32</td>
</tr>
<tr>
<td>11</td>
<td>Peoples R China</td>
<td>3</td>
<td>Australia</td>
<td>18</td>
<td>Italy</td>
<td>29</td>
</tr>
<tr>
<td>12</td>
<td>Italy</td>
<td>2</td>
<td>Italy</td>
<td>13</td>
<td>Netherlands</td>
<td>28</td>
</tr>
<tr>
<td>13</td>
<td>Wales</td>
<td>2</td>
<td>Denmark</td>
<td>12</td>
<td>Brazil</td>
<td>26</td>
</tr>
<tr>
<td>14</td>
<td>Portugal</td>
<td>2</td>
<td>South Korea</td>
<td>12</td>
<td>Taiwan</td>
<td>24</td>
</tr>
<tr>
<td>15</td>
<td>South Africa</td>
<td>2</td>
<td>Scotland</td>
<td>11</td>
<td>Sweden</td>
<td>23</td>
</tr>
<tr>
<td>16</td>
<td>Sweden</td>
<td>2</td>
<td>South Africa</td>
<td>10</td>
<td>Switzerland</td>
<td>21</td>
</tr>
<tr>
<td>17</td>
<td>Austria</td>
<td>1</td>
<td>Taiwan</td>
<td>10</td>
<td>Belgium</td>
<td>17</td>
</tr>
<tr>
<td>18</td>
<td>Switzerland</td>
<td>1</td>
<td>India</td>
<td>10</td>
<td>India</td>
<td>16</td>
</tr>
<tr>
<td>19</td>
<td>Hungary</td>
<td>1</td>
<td>Sweden</td>
<td>10</td>
<td>Norway</td>
<td>14</td>
</tr>
<tr>
<td>20</td>
<td>Tunisia</td>
<td>1</td>
<td>Switzerland</td>
<td>9</td>
<td>Iran</td>
<td>14</td>
</tr>
</tbody>
</table>

As shown in Table 5, the number of publications by the United States ranked first in all the 3 categories. German ranks 2nd, 3rd, 3rd in “remarkably followed”, “fairly followed” and “poorly followed” papers respectively. China ranks 11th in “remarkably followed” papers, while ranks 2nd in both “fairly followed” and “poorly followed” papers. Both the author analysis and country/region analysis indicate that the medical
research in the United States got the highest degree of public interest, followed by German. The medical research in China got a relatively low degree of public interest, which means that the quality of medical research in China need to be furtherly improved.

Paper length analysis

The average page count for “remarkably followed”, “fairly followed” and “poorly followed” papers are 18.21, 17.38, and 15.75 respectively (Table 4). This means that papers with higher degree of public interest always have more abundant content than those with lower degree of public interest.

Reference analysis

The average reference count for “remarkably followed”, “fairly followed” and “poorly followed” papers are 49.48, 51.15, and 44.21 respectively (Table 6). This means that papers with higher degree of public interest always cite more sources.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Number of Pages (Mean ±SD)</th>
<th>Number of References (Mean ±SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Remarkably Followed&quot; Papers</td>
<td>18.21 ± 4.06</td>
<td>49.48 ± 8.67</td>
</tr>
<tr>
<td>&quot;Fairly Followed&quot; Papers</td>
<td>17.38 ± 3.07</td>
<td>51.15 ± 7.31</td>
</tr>
<tr>
<td>&quot;Poorly Followed&quot; Papers</td>
<td>15.75 ± 2.57</td>
<td>44.21 ± 5.12</td>
</tr>
</tbody>
</table>

Discussion

In terms of Twitter, Mendeley and Counter, non-medical papers received more attentions than medical ones, but in term of PMC, the page views and downloads for medical papers on PubMed Central were greater than non-medical ones. This may be in part because of the nature of medical research. The information provided by medical paper is hard to understand for those without a medical background, so the dissemination of medical knowledge are restricted within the domain of medical scientists other than the public. For both medical and non-medical papers, Twitter, Mendeley and Counter showed a sharp increase over time at the very beginning after the papers were published. Then, the increase gradually slowed down. This finding was further confirmed by the fact that 90.87% medical papers’ Counter increase reached the peak within a week. One of the best ways to strengthen the medical research infrastructure is by promoting it to the public. So the medical research community should take advantage of the social media to make medical information to be disseminated in a way that is easy to understand for those without a medical background.

We grouped all the medical papers into 3 categories with Counter’s Characteristic Score, that is, “remarkably followed” papers, “fairly followed” papers and “poorly
followed” papers. We compared the most frequent words in the abstracts for the 3 categories, the results shows that cancer research is still the biggest concern for the medical scientists. Both the author analysis and country analysis shows that medical research in China got a relatively low degree of public interest compared to the United States and German. So medical scientists in China should attach great importance to quality of their research. The results of paper length analysis and reference analysis indicate that papers with more content and references tends to have a higher degree of public interest. This provide with the medical scientists a good way to prepare their research papers.

**Conclusion**

In conclusion, this study investigated the diffusion pattern of medical knowledge in social media by using PLOS ALM data. The data provided are informative for the scientists in medical field as well as policy makers and administrators for managing and financing medical research in the future.

**Acknowledgments**

The research reported in this paper was done as part of the “Theoretical and Empirical Study of Micro-level Entity Evaluation in Scientific Documents through Multiple Analyses - An Empirical Study of Biomedical Science” (No. 71573162), supported by National Natural Science Foundation of China. And it was also supported by Program for the Top Young Academic Leaders of Higher Learning Institutions of Shanxi, China.

**References**


TYPE: Full Paper

An Integrated Analysis of Topical and Emotional Evolution of Microblog Public Opinions on Public Emergencies

An Lu¹ Wu Lin² Chuanming Yu³,*

¹anlu97@163.com
Wuhan University, Wuhan (China)

²2016201040013@whu.edu.cn
Wuhan University, Wuhan (China)

³yucm@zuel.edu.cn
Zhongnan University of Economics and Law, Wuhan (China)
*corresponding author

Abstract
Microblog is an important medium of communicating public opinions toward public emergencies. Excavating topics and sentiment of microblogs on public emergencies has important practical significance in grasping online public opinions, identifying and predicting potential problems and risks during public emergencies. In this study, we propose an approach to analyse the topical and emotional evolution of microblog public opinions on public emergencies. The Zika outbreak is taken as an example and the life cycle of related microblog public opinions is divided into several phases. Topics are extracted from microblogs by the word2vec technique. The sentiment analysis has been conducted based on dictionaries containing sentiment words and emoticons to categorize the sentiment of microblogs at a fine-grained level. The emotional intensity of microblogs for each topic is also calculated to achieve synergetic analysis of topics and sentiment of microblogs. The proposed method can reveal the topical and emotional features and emotional intensity of microblogs on public emergencies and summarize the synergetic evolution patterns of topics and sentiment of online public opinions.

Conference Topic
Knowledge discovery and data mining

Introduction
With the development of the Internet, microblog, WeChat, and other social media are changing people's way of information dissemination and consumption to a great extent. The functions of newspapers, television, radio and other traditional media have been gradually replaced by social media. Sina Weibo is one of the most important social media platforms in China. As of September 30th, 2016, Sina Weibo has owned 297 million monthly active users and 132 million daily active users, among whom 73.9% of the users are concerned about the real-time news and hot topics (Fan, 2017). Users can freely publish posts and enterprises, institutions and government departments set up official microblog accounts to release authoritative information. Thus, microblog has been an important channel of current public opinions.

As public emergencies have the characteristics of suddenness and uncertainty, online public opinions also have similar features and need to be processed timely. The excavation and analysis of the topics and corresponding emotions of microblog entries about public emergencies can help emergency administrators understand the public's attitude toward a certain event and related policies. Thus, the government, enterprises and other organizations can actively respond to the emergency, effectively ease the social negative emotions, and promote social harmony.
and stability. In this study, we aim to construct the method of analyzing microblog public opinion evolution patterns. The method integrates the topical and emotional characteristics, reveals the synergistic pattern between the topics and emotions of microblog public opinions and provides a scientific and rational basis for decision-making and risk prediction by emergency management departments. The purpose of this study is 1) to construct an updated emoticon dictionary and combine the emoticons with short texts to capture the micro-blog’s hidden emotions accurately; 2) to extend the affective tendency analysis from positive, neutral and negative attitudes to a more fine-grained emotion classification; and 3) to capture topics with their related emotions at each stage, analyze and visualize their evolution pattern.

Related research

Topical analysis of microblogs
In recent years, researchers have paid much attention to the analysis of the topical evolution of microblogs, in which the Latent Dirichlet Allocation (LDA) model is frequently used. The LDA model considers a topic as a set of flat probabilities and does not take into account the position relationship between words. The resulting topics are isolated from each other. Thus, it often misses some key features, e.g. the position correlation between topics. In addition, the effectiveness of the LDA model is largely influenced by the length of the document. Microblog entries are often very short with scattered contents and formats. Thus, the performance of the traditional LDA model on microblog is often very poor.

In recent years, with the breakthrough application of deep learning in the field of natural language processing, more and more studies have shown that the word2vec model (Mikolov et al, 2013) which is based on the deep learning thought is significantly better than the LDA model. Word2vec expresses words in word vectors, and the similarity among the word vectors is used to measure the similarity among the texts. It can overcome the deficiency of the bag of words and mine the association among words to obtain abundant semantic information. Studies that apply the word2vec model to analyzing and clustering topics are emerging. A number of studies have shown that the word2vec model is more suitable for extracting topics from short texts, such as microblog entries than the LDA model (Zhu, 2014; Li et al., 2015; Chen et al., 2015).

Sentiment analysis of microblogs
Sentiment analysis includes subjective and objective analyses of emotions, sentiment polarity analysis and sentiment classification. At present, related studies often classify emotions into positive, negative and neutral emotions. However, Internet users’ emotional attitudes in microblogs are often not of a simple type and need more fine-grained sentiment classification instead, such as pleasure, anger, sorrow and joy. This kind of studies is at their early stage. Due to the correlation and combination of various emotions, multiple sentiment classification is regarded as one of the most complicated tasks (He et al, 2017).

The commonly used sentiment analysis methods include the sentiment dictionary-based approach and supervised machine learning (Tang et al., 2016). Sentiment dictionaries can be used to extract subjective vocabularies and their emotional tendencies in different contexts with high accuracies. As supervised machine learning methods depend on labeled training sets and have limited applicability in different fields (Mikolov et al, 2013), in this study we employ the sentiment analysis method based on the sentiment dictionary.
In the traditional text analysis, pictures and emoticons are often seen as noise data. However, in the sentiment analysis of microblog entries, they are likely to become an important source of emotional expression. Some researchers have studied the purpose of use, characteristics, effect and distribution of emoticons. However, most of the existing emoticons dictionaries are labelled based on their official definitions by microblog platforms. Actually, with the development of network cultural, the meanings of emoticons become gradually different from their original definitions. Therefore, in this study we construct an updated emoticon dictionary which is synchronized with the current network culture.

**Time series analysis of microblog public opinions**

The evolution process of public emergencies has a certain life cycle. Extensive research has been done on the spreading process of public opinions. These studies divide public emergencies into several stages based on the sequence of events, such as the three-phase model, the four-stage model (Robert, 2004). A recent study by Jia et al. (2015) divides the life cycle of public emergencies into the initial stage, the outbreak stage, the decline stage and the settling stage. Combined with the characteristics of public emergencies, in this study we adopt the life cycle division method by Jia et al. (2015).

In this study, we construct a topical and emotional evolution analysis framework to mine the microblog public opinions on public emergencies. The topics in microblog entries are extracted by the method based on word2vec and sentiment analysis is conducted based on the constructed emoticon dictionary, in which emotional words, emoticons and other sentiment source are introduced. Fine-grained emotions are identified for each topic and the corresponding emotional intensities are calculated. Combined with the life cycle of the public emergency, the hidden topics and emotions of microblog entries are detected simultaneously.

**Research framework and methods**

**Research framework**

To construct the evolution model of topics and emotions of microblog public opinions on public emergencies, the overall research framework is shown in Figure 1, which consists of three procedures, i.e. life cycle division, topic clustering and sentiment analysis. First, we crawl the microblog entries together with their comments on a certain public emergency, e.g. the Zika outbreak, and preprocess the microblog corpus. As different public emergencies may have different number of peaks, their life cycle of public opinions is divided accordingly. The evolution period of public opinions is shown as follows.

\{\text{the initial stage, the outbreak stage, the decline stage, the settling stage}\}, N=1
\{\text{the initial stage, the outbreak stage, the first recession stage, the second growth stage, the second recession stage, the settling stage}\}, N=2
\{\text{the initial stage, the outbreak stage, the first recession stage, the second growth stage… the N-1}^{\text{th}} \text{ recession stage, the N}^{\text{th}} \text{ growth stage, the N}^{\text{th}} \text{ recession stage, the settling stage}\}, N>2

where N is the number of peaks for public opinions. When N equals 1, the public emergency is defined as a unimodal event. When N equals 2, it is defined as a bimodal event. When N is larger than 2, it is defined as a multimodal event.

Second, the word2vec model is used to cluster the topics in the microblog corpus of each stage. Third, the mutual information method is employed to construct emoticon dictionary, which is combined with the emotion dictionary by Dalian University of Technology (Xu et al.,
We classify the comments of the blogs and calculate the emotional intensity. Finally, we get the trend of micro-blogging theme and emotion of different stages of the life cycle and visualize it.

**Figure 1 Research procedures of topical and emotional evolution of microblogging public opinion**

*Research methods and procedures*

*Theme extraction based on word2vec*

The word2vec model is a tool that converts words into word vectors proposed by Mikolov et al. (2013) which belongs to one of the shallow neural network language models. Word2vec can be used to map the processing of text content to vector operations in K-dimensional vector space, so the similarity of text semantics can be expressed by the similarity of space vector. Word2vec uses CBOW and Skip-gram to construct the vector representation of words, the CBOW model predicts the probability of the current term based on the context, while Skip-Gram is the opposite: the probability of predicting the context according to the current term. This paper is based on the Skip-gram model, using Python's genism (Řehůřek et al., 2011) module to provide the Word2Vec toolkit for training.

In this paper, we use the method of word2vec and K-means clustering to analyze the blog data set of each stage of the evolution of public opinion. Word2vec will express the words in the blog into a word vector containing semantic relations based on blog data set, while K-means uses the word vector as the clustering data sets, and Euclidean distance standardized by Z-score
will be used to calculate word similarity between two words, namely the smaller the European distance, the greater the similarity. In this way, the theme of the evolution of public opinion in each stage is obtained by clustering results.

The construction of the emoticons dictionary

This paper refers to the Dalian University of Technology’s affective lexicon ontology library (Xu et al. 2008) (happy good, anger, sadness and fear, evil, surprise) to classify the emoticons. According to the characteristics of the microblog of the public health emergencies, we combine "happy" and "good" to "happy", "anger" and "evil" to "anger", and finally we get five categories of emotion: "happy", "anger", "sadness", "fear", "surprise" and classify the emoticons. Microblog’s comments are short, different emoticons in the same emotion expression tend to have strong association (Li et al., 2016), therefore, the feelings expressed by emoticons in one comment always have a positive correlation, so the emoticons dictionary is constructed by mutual information calculated method. Based on the probability of co-occurrence of two different emoticons, a certain category of emoticons (seed emoticons) is used to determine the emotional categories of the other emoticons to classify emotions.

According to Ma et al. (2016), microblog emoticons have a long tail effect, the frequency of emoticons in the top 10% can cover the most of the emotion. In order to generate a seed emoticons table, we marked the categories and intensities of these top 10% high frequency emoticons. The results are shown in table 1.

Table 1 Seed emoticons (part)

<table>
<thead>
<tr>
<th>emoticons</th>
<th>category</th>
<th>intensity</th>
<th>emoticons</th>
<th>category</th>
<th>intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>[doge]</td>
<td>surprise</td>
<td>5</td>
<td>[心]</td>
<td>happy</td>
<td>5</td>
</tr>
<tr>
<td>[笑 cry]</td>
<td>happy</td>
<td>9</td>
<td>[挖鼻]</td>
<td>surprise</td>
<td>3</td>
</tr>
<tr>
<td>[哈哈]</td>
<td>happy</td>
<td>7</td>
<td>[偷笑]</td>
<td>happy</td>
<td>5</td>
</tr>
<tr>
<td>[微笑]</td>
<td>anger</td>
<td>5</td>
<td>[哀]</td>
<td>fear</td>
<td>7</td>
</tr>
<tr>
<td>[赞]</td>
<td>happy</td>
<td>7</td>
<td>[生病]</td>
<td>sadness</td>
<td>5</td>
</tr>
</tbody>
</table>

Note: The emotional intensity is divided into 1, 3, 5, 7, 9, 9 for the maximum strength, 1 for the minimum strength.

Microblog integrated affective computing

In this study, we analyze the comments of the microblogs, select the emotional words, emoticons, commonly used adverbs of the Internet (Han et al., 2012), and extract the negative words from HowNet (CNKI, 2017) to calculate the comprehensive emotions of the comments.

The relationship between the emotional intensity of the emoticons and the emotional intensity of the emotional word will directly affect the final emotional trend analysis. This paper holds that the emotions expressed by the emoticons are as important as the emotional words. Since the annotation (category and intensity) of the emoticons refers to the annotation system of the text emotion dictionary, so the same type of emoticons and words can be calculated together.

Traditional sentiment analysis methods only consider the polarities of emotions. Negative words tend to reverse the polarities. However, the emotions in this study are divided into five
types so that the treatment of negative words is more complicated. We refer to the negative lexicon proposed by Du. (2013), which is supplemented by the corpus of this study, and forms the subdivided emotional conversion table after the negative word modification, as shown in Table 2. The emotional correction formula of the emotional word with an odd number of negative words is shown in equation (1). When there are even negative words to modify the emotional word, the emotional word is not adjusted.

$$\Psi(\omega_i) = \frac{1}{2} \sqrt{S(\omega_i)} \quad (1)$$

Where \(\Psi(\omega_i)\) represents the emotional word intensity after the negative word modification, and \(S(\omega_i)\) is the initial emotional intensity of the affective word \(\omega_i\).

**Table 2 Conversion of emotional category after negative words’ modification**

<table>
<thead>
<tr>
<th>Original emotional category</th>
<th>Negative modified emotional category</th>
</tr>
</thead>
<tbody>
<tr>
<td>happy</td>
<td>sadness</td>
</tr>
<tr>
<td>anger</td>
<td>--</td>
</tr>
<tr>
<td>sadness</td>
<td>a*happy</td>
</tr>
<tr>
<td>fear</td>
<td>--</td>
</tr>
<tr>
<td>surprise</td>
<td>--</td>
</tr>
</tbody>
</table>

Table 2

Note: A is the real number of \([0,1]\), which indicates that the intensity of emotion is weakened, and the accuracy of emotion recognition is highest when measuring \(a=0.1\).

In addition, the degree of adverbs have different degree of adjustment, the emotional word will be modified by the degree of adverbs and obtains a new emotional intensity value, \(Value_{adv}\) represents the adjustment of adverbs around emotional words \(\omega_i\), the calculation is shown in equation (2). According to the statistical comparison, it is found that the degree adverbs and negative words in the range of no more than three words can be recalled well. Therefore, the sliding window value is set to 3, and the emotional word and the corresponding degree adverb, Negative words are matched.

$$\Phi(\omega_i) = Value_{adv} \cdot s(\omega_i) \quad (2)$$

In one comment, the emotional words and emoticons of the same emotional category are summed up, and the emotion categories are \(p, p \in \{\text{happy, anger, sadness, fear, surprise}\}\). The emotional calculation of the comments based on the emotional words and emoticons are shown in equation (3).

$$Value(E_p) = \sum_{i=0}^{n}[f(\omega_i) + g(e_i)] \quad (3)$$

Where \(Value (E_p)\) represents the emotional value of the comment, \(g(e_i)\) represents the emotional value of the emoticons, the final emotional category of the comment and the emotional value are determined by the most significant value of the emotional category and its corresponding value, \(n\) is the sum of the same kind of emotional features (emotional word and emoticons) in the same comment.

**An Analysis of theme and Emotional Evolution of Microblogging Public Health Emergencies**

**Data collection and preprocessing**

We set “Zika” as key word to retrieval blogs from Sina microblog platform. The data set is all relevant microblog and its corresponding comments (direct comment and Forward the...
comments) from "2016-1-1" to "2016-12-31", After data cleaning such as removing the link in the text, @ other users and the stop words, we get 108,997 data.

4.2 Data analysis
We divide the data of Zika microblog as the method provided by Jia et al. (2015). As shown in Figure 2.

![Figure 2 The life cycle of Zika microblog spreading](image)

In this paper, we use word2vec to summarize the theme of different stages of topics, and summarize the theme of each stage as shown in Table 3.

<table>
<thead>
<tr>
<th>Cycle stage</th>
<th>Subject number</th>
<th>Summary of topics</th>
<th>Cycle stage</th>
<th>Subject number</th>
<th>Summary of topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>the initial stage</td>
<td>Topic I-1</td>
<td>Multiple cases of Zika in Europe</td>
<td>the second growth stage</td>
<td>Topic IV-1</td>
<td>Mosquito sterilization against the Zika</td>
</tr>
<tr>
<td></td>
<td>Topic I-2</td>
<td>The sudden increase in neonatal small head syndrome in the Americas</td>
<td></td>
<td>Topic IV-2</td>
<td>Rio Olympic Games affected by Zika</td>
</tr>
<tr>
<td></td>
<td>Topic I-3</td>
<td>Brazilian carnival and the Olympics are affected</td>
<td></td>
<td>Topic IV-3</td>
<td>Singapore and other countries found a series of Zika virus cases</td>
</tr>
<tr>
<td></td>
<td>Topic I-4</td>
<td>WHO announced the establishment of the emergency committee in Geneva</td>
<td></td>
<td>Topic IV-4</td>
<td>Chinese mosquito nets and Microsoft smart mosquitoes became popular in the Rio Olympic Games</td>
</tr>
<tr>
<td>Topic II-1</td>
<td>The case of imported cases of infection of the Zika in many provinces of China</td>
<td>the second recession stage</td>
<td>Topic V-1</td>
<td>Henan confirmed the first case of Zika</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>Topic II-2</td>
<td>China found the third case of Zika</td>
<td></td>
<td>Topic V-2</td>
<td>The outbreak of Zika in Southeast Asia</td>
<td></td>
</tr>
<tr>
<td>Topic II-3</td>
<td>China imported Zika infection patient is still receiving isolation therapy</td>
<td></td>
<td>Topic V-3</td>
<td>There is a risk of Zika infection at home and abroad</td>
<td></td>
</tr>
<tr>
<td>Topic II-4</td>
<td>Two cases of imported cases of Zika infection in Guangdong</td>
<td></td>
<td>Topic V-4</td>
<td>The way of transmission and prevention of Zika in holiday travel</td>
<td></td>
</tr>
<tr>
<td>Topic II-5</td>
<td>Mainland China diagnosed the first case of imported Zika infection</td>
<td>the settling stage</td>
<td>Topic VI-1</td>
<td>A new breakthrough in the research of domestic and foreign Zika vaccine</td>
<td></td>
</tr>
<tr>
<td>Topic II-6</td>
<td>China completes genome sequence determination of Zika</td>
<td></td>
<td>Topic VI-2</td>
<td>Recent case reports and studies of Zika</td>
<td></td>
</tr>
<tr>
<td>Topic III-1</td>
<td>The occurrence of imported cases of Zika in many provinces of China</td>
<td></td>
<td>Topic VI-3</td>
<td>Prevention and control of Zika diseases in autumn and winter</td>
<td></td>
</tr>
<tr>
<td>Topic III-2</td>
<td>Many areas in China carry out the Patriotic Health Campaign in spring</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic III-3</td>
<td>Beijing, Guangdong and more provinces found Zika virus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic III-4</td>
<td>Worried about the Zika virus, McIlroy gave up the Rio Olympics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic III-5</td>
<td>Beijing and Guangdong emerged the new case of imported Zika virus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
According to the law of the time sequence of the public opinion propagation, we carry on the sentiment analysis of the corresponding comment of microblog in the life cycle. As shown in figure 3. From the public opinion of the evolution of the entire life cycle, the initial stage and the outbreak stage of public opinion mainly led by the "fear" and "anger", indicating that public opinion has been the phenomenon of emotional agglomeration. The intensity of the "anger" in the initial stage has reached the highest value of this emotional category throughout the evolutionary period, which means the reports of the epidemic abroad have aroused the intense boredom and hatred of the domestic netizens (TopicI-1, TopicI-2, TopicI-4). In the stage of the outbreak, the intensity of "fear" is the highest value of this emotional category in the entire evolutionary cycle, this also means that the epidemic spread from abroad to China (TopicII-5), internet users' emotions have been changed from "anger" to "fear", netizens’s position is gradually converted from the outsider to the party. However, in the first recession, the intensity of the emotions was lower than the previous stage, the proportion of emotional composition is relatively uniform, "anger" "fear" "happy" three emotions cross, the situation of public opinion has eased. There are two main reasons for this phenomenon: On the one hand, the spread of the epidemic in the country caused panic and anger, on the other hand is due to the patriotic health movement throughout the country, various measures are conducted to prevent the spread of negative emotions, which means emergency management has been effective. In the second growth stage, "anger" and "fear", as negatively dominant mood in the previous stage, have been greatly reduced, "happy" has become the dominant emotion, the wind of public opinion began to reverse. Due to the hosting of the Rio Olympics, news media’s reports have shifted the attention from epidemic to the sports event. In the second recession, "fear" compared to the previous stage began to rise, due to the fact that Henan province reported the first Zika(TopicV-1) in the province and Southeast Asia begins to erupt Zika(TopicV-2), the situation of crisis management has intensified again. In the subsidence stage, "happy" and “sad" become the dominant categories of emotion, the reason is that domestic and foreign research has made a breakthrough in the development of Zika vaccine, the majority of people maintain an optimistic attitude toward the treatment of Zika. However, it can not be ignored that the intensity of the "sad" in this stage has reached the highest value of this emotional category throughout the evolutionary period, so the conciliation after the epidemic is of critical importance.

Therefore, the analysis of the emotional evolution of public emergency at different stages has brought a new perspective for crisis management. Throughout the evolutionary cycle, the crisis management department should provide accurate information on the cause and reasonable consequences of the incident in the initial stage to resolve the confusion of Internet users, curb the fermentation of rumors, avoid causing a large area of fear. Emergency management departments should promptly analyze the composition of sentiment of Internet users and
observe the fluctuations of the emotional value to look for the reasons of the fluctuations and carry out targeted emotional guidance and intervention through various media reports, so as to avoid or reduce the impact of public opinion when the event broke out; When the event outbreak, the emergency management departments should promptly report on the direction of the emergencies and assess the safety risk of the incident and popularize precautionary measures. The media should follow up the latest measures and responses at home and abroad in real time to lead Internet users to move toward positive emotions; In the first recession and the second outbreak, crisis management department still cannot lower the guard and should monitor the themes and proportion of emotions to prevent the recurrence of public opinion; In the settling stage, crisis management department should carry out psychological aid and post-disaster reconstruction to placate the public mood. In addition, the "anger" has the meaning of anger and censure, it can not only embody public’s attitudes to public emergency, but also reflect the feedback of crisis management of the relevant departments. If the mood continues to be high, it is time to look at themselves and see if the timeliness and effectiveness of crisis management need to be strengthened.

![Figure 3 Emotional types and intensities in different stages](image)

**Conclusion**

In this paper, we proposed a topic and emotion fusion evolution analysis method, taking deep learning into the theme analysis of public opinion in microblog. Based on the word2vec model, the microblog’s content of the each stage of is modeled. The word in the microblog is expressed as word vector which contains the semantic relations, and the theme of the each stage in evolution of public opinion can be obtained by K-means clustering; We adopt more fine-grained sentiment classification and rich emoticons from existing dictionaries, we also make sentiment analysis for the themes in each stage based on emotional words, emoticons and other multiple source of emotion and obtain the map which shows the evolution trend of topic and emotion fusion evolution. A total of 108,997 samples were collected from the blogs and their corresponding comments, which shows that the method can effectively obtain the theme and the emotion trend in the evolution of public opinion. The research results provide reference for the emergency management departments and other stakeholders to obtain real-time public opinion in terms of cognition, attitude, emotion and behavioral tendency of the internet users.
It is also helpful for the related organizations to conduct the accurate prediction and risk control of the network public opinion, and provides efficient information service and emergency management.

Acknowledgments
This study is funded by the Natural Science Foundation of China (No. 71603189 and 71373286) and Youth Innovation Corps Fund of Humanities and Social Sciences, Wuhan University.

Reference
Journal of Hebei University, Natural Science Edition, 35 (6), 651-656.
Communicating Scientific Video Articles on Twitter: An Initial Exploration of JoVE Publications

Shenmeng Xu 1 Houqiang Yu 2 Bradley M. Hemminger 3 Xie Dong 4

1 shenmeng@email.unc.edu
University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (United States)

2 yuhouq@yeah.net
Nanjing University of Science and Technology, Nanjing, Jiangsu (China)

3 bmh@ils.unc.edu
University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (United States)

4 xied@email.unc.edu
University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (United States)

Abstract
This paper investigates how and why scientific video articles are communicated on Twitter. Specifically, we study video articles published in the Journal of Visualized Experiments (JoVE). We harvested all of the tweets from October 2011 to November 2015 that contain one or more JoVE links. These tweets “citing” JoVE articles were analyzed both statistically and qualitatively. In this research-in-progress paper, we discuss our preliminary findings on the distribution of these tweets, with a brief description of the affordance use of Twitter such as hashtags and mentions. The content analysis of the tweets is still in progress. We present the coding schemes of both the Twitter user accounts and the tweets’ text formed after two rounds of coding, discussion, and revision. We also briefly discuss issues including the video/visual feature mentioned, the role of Twitter bots, and self-promotion of different stakeholders in the Twitter communication of JoVE video publications.

Conference Topic
Altmetrics
Science Communication
Participation in science
Twitter

Introduction
Research and education are increasingly relying on a more diverse array of information sources. Academic uses of videos both in education applications (Jones & Cuthrell, 2011) and for research purposes (Kousha et al., 2012) have been studied. Previous studies have revealed potential benefits that videos could bring to education in the arts and humanities (Brook, 2011; Cayar, 2011; Hoskins, 2009; Muniandy & Veloo, 2011; Rees, 2008); meanwhile, research has demonstrated considerable advantages of using videos in science and medicine education (Clifton & Mann, 2011; Fernandez et al., 2011; Franz, 2012; Kaw & Garapati, 2011; Knösel et. al, 2011; Settle et al., 2011). The pedagogical value of online videos could help enhance researchers’ knowledge and skills, thus promoting the accuracy and efficiency of their work. Therefore, we are particularly interested in how videos are being utilized in the scientific research process.

Previous studies have been mostly focused on Youtube videos as scholarly information sources; Here we investigate a peer reviewed scientific journal specifically publishing curated videos, the Journal of Visualized Experiments (JoVE). JoVE is a 10-year old journal with nearly 7,000 publications to date. Professional videographers and editors assist in producing videos to visualize methods, data analysis and results in both the physical and life sciences. According to
JoVE’s website (JoVE, 2017), it is devoted to helping researchers understand complicated experimental techniques and methods to make them easier to comprehend and reproduce. Pasquali (2007) discussed the benefits of using videos to communicate scientific protocols instead of using verbal descriptions. He emphasizes the important role that human’s visual perception plays in comprehending ideas and feelings, and argues that because a video includes “information such as color, position, duration, shape, and motion,” it is the “optimal format for transmitting the manifold details of new protocols or technical procedures.”

The microblogging service Twitter is one of the most important altmetrics data sources. In recent years, academic activity on Twitter has been extensively studied (Priem & Costello, 2010; Thelwall et al., 2013; Holmberg & Thelwall, 2014; Holmberg et al., 2014; Haustein et al., 2016; Eriksson-Backa, Holmberg & Ek, 2016; Vainio & Holmberg, 2017). This exploratory research-in-progress paper is particularly focused on the communication of JoVE articles on Twitter and is driven by the following basic research questions:

1. Who tweets JoVE articles? (account identify, account gender, and Twitter bots’ role)
2. How do they tweet JoVE articles? (elements included in the construction of tweets, sentiment, and affordance use such as hashtags and mentions)
3. Why do they tweet JoVE articles? (motivation of tweeting, with a special focus on different stakeholders’ self-promotion)
4. Are video/visual elements specifically emphasized in tweets of JoVE articles?

Methods and Data

Altmetric.com provided the full dataset from October 2011 to November 2015 for this research. This dataset contains over 4.4 million records in JSON format. Each record represents one scholarly artifact and encompasses various types of altmetrics data about this artifact. Specifically, these altmetrics include those based on Twitter, Facebook, Blogs, Wikipedia, News, Google Plus, Policies, Reddit, F1000, Peer Reviews, Videos, and Q&As. After learning the structure of this dataset, we established a relational database. In addition, we used a Python script to process the data and were able to successfully extract all of the data of tweets that contain links to JoVE publications.

Overall, 7,500 tweets were extracted, with data about their user names and URLs, as well as the titles and URLs of the JoVE publications they contain. To better understand the content of the tweets, we used another Python script to retrieve tweets’ text from Twitter.com. Considering that the altmetrics data were gathered earlier than we started to work on this project (in September 2016), some tweets were shown to had been deleted. As a result, we further cleaned the data, and excluded 308 “HTTP Error 404: Not Found” ones and 798 “Twitter/Account Suspended” ones. At the end, when we were exploring tweets content related questions, our data consisted of 6,394 tweets. These data were processed one step further using a third Python script, to extract the hashtags and mentions in the tweets’ text.

In order to gain a deeper understanding of the construction patterns of tweets and the motivation behind them, content analysis was conducted by three authors. Considering that one user can post multiple tweets, we coded the Twitter user accounts and the tweets’ text separately. In each process, there were multiple steps. Among the 6,394 tweets, we planned to use stratified sampling based on the activeness of users (the number of JoVE article tweets they created). For coding the tweets, we firstly studied relevant literature. 30 tweets were then coded by three coders. The initial coding scheme was formed after our discussion about the first coding process. In the second step, 90 tweets were coded. Reaching an average agreement rate of 70%, three coders discussed again, with a particular focus on the coding disagreement and potential new categories. The final coding scheme of tweets was formed after that. In the following step, two coders will code 480 more tweets. In addition, we went through a similar process of coding the accounts. An agreement of 76% was reached and the final coding scheme of accounts was
formed. In the following step, two coders will code 185 more accounts. The third coder will review the disagreement if needed. Due to limited space and the work-in-progress nature of this article, we will present our coding schemes in the Results section.

Results

Distribution of Tweets

In total, 2,153 Twitter users have posted 7,500 tweets that contain JoVE article links. Figure 1 shows the skewed power-law distribution of tweets created by users. Specifically, the top 9 most active users (0.42%) have tweeted more than half of the tweets. By contrast, 1,676 users (77.84%) have only tweeted one JoVE article each. Almost 90% (89.13%) of users have either tweeted one or two JoVE articles.

More details of the top 10 most active Twitter accounts are shown in Table 1. According to our coding and discussion, the top 10 most active Twitter accounts include at least six bot accounts. The bots could be impersonated in institution accounts or individual accounts. The rest four accounts are also likely to be combining bot tweeting techniques with manual tweeting.

![Figure 1. Distribution of tweets created by Twitter users](image)

Table 1. Top 10 most active Twitter accounts in tweeting JoVE articles

<table>
<thead>
<tr>
<th>Account name</th>
<th>No.</th>
<th>%</th>
<th>Cumulative %</th>
<th>Coded Account Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>JoVEJournal</td>
<td>1506</td>
<td>20.08%</td>
<td>20.08%</td>
<td>JoVE the journal</td>
</tr>
<tr>
<td>postechlibrary</td>
<td>800</td>
<td>10.67%</td>
<td>30.75%</td>
<td>digital library</td>
</tr>
<tr>
<td>SfakianakisAI</td>
<td>394</td>
<td>5.25%</td>
<td>36.00%</td>
<td>medical institution (bot)</td>
</tr>
<tr>
<td>Sfakianakismed</td>
<td>384</td>
<td>5.12%</td>
<td>41.12%</td>
<td>medical institution (bot)</td>
</tr>
<tr>
<td>alsfakiahotmail</td>
<td>375</td>
<td>5.00%</td>
<td>46.12%</td>
<td>medical institution (bot)</td>
</tr>
<tr>
<td>rnomics</td>
<td>121</td>
<td>1.61%</td>
<td>47.73%</td>
<td>RNA researcher (bot)</td>
</tr>
<tr>
<td>fly_papers</td>
<td>68</td>
<td>0.91%</td>
<td>48.64%</td>
<td>paper feed (bot)</td>
</tr>
<tr>
<td>MPritsker</td>
<td>58</td>
<td>0.77%</td>
<td>49.41%</td>
<td>CEO, co-founder of JoVE</td>
</tr>
<tr>
<td>Medgadget</td>
<td>49</td>
<td>0.65%</td>
<td>50.07%</td>
<td>medical technology news feed</td>
</tr>
<tr>
<td>semantic_bot</td>
<td>46</td>
<td>0.61%</td>
<td>50.68%</td>
<td>paper feed (bot)</td>
</tr>
</tbody>
</table>
Most Frequently Used Hashtags and Mentioned Users

In total, there are 3,235 hashtags (#) tweeted in all of these 6,394 JoVE article tweets. In other words, approximately one hashtag is used in every two tweets. Among these hashtags, there are 928 unique ones, with the most frequently used one tweeted 862 times. We coded the top 100 frequently used hashtags into six categories in terms of their purpose of use: 1) to indicate that this tweet is about an article (which contains more information than the 140 words limit), 2) to indicate that this tweet is about JoVE, 3) to emphasize the visual/video feature of JoVE articles, 4) to emphasize the technical feature of JoVE articles, 5) to specify relevant research areas/topics, and 6) to specify relevant experimental techniques/tools/objects. Due to limited space here, we present the top 10 in Table 2 below. With #article, #Jove, #scivideo (scientific video), and #MySciVid (my scientific video) being the top four, the rest six all specify relevant research areas/topics. The next step of our study is to code all of the hashtags to better understand the use of hashtags.

Table 2. Top 10 most frequently used hashtags

<table>
<thead>
<tr>
<th>Hashtag</th>
<th>No.</th>
<th>Coded Purpose of Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>#article</td>
<td>862</td>
<td>to indicate the content of the link</td>
</tr>
<tr>
<td>#Jove</td>
<td>188</td>
<td>to indicate that this tweet is about JoVE</td>
</tr>
<tr>
<td>#scivideo</td>
<td>139</td>
<td>to emphasize the visual/video feature of JoVE articles</td>
</tr>
<tr>
<td>#MySciVid</td>
<td>60</td>
<td>to emphasize the visual/video feature of JoVE articles</td>
</tr>
<tr>
<td>#Genetics</td>
<td>59</td>
<td>to specify relevant research areas/topics</td>
</tr>
<tr>
<td>#RNA</td>
<td>58</td>
<td>to specify relevant research areas/topics</td>
</tr>
<tr>
<td>#Cancer</td>
<td>56</td>
<td>to specify relevant research areas/topics</td>
</tr>
<tr>
<td>#neuroscience</td>
<td>56</td>
<td>to specify relevant research areas/topics</td>
</tr>
<tr>
<td>#Immunology</td>
<td>52</td>
<td>to specify relevant research areas/topics</td>
</tr>
<tr>
<td>#microbiology</td>
<td>40</td>
<td>to specify relevant research areas/topics</td>
</tr>
</tbody>
</table>

In terms of mentions (@), all these tweets contain 1457 mentions of 483 unique accounts. Due to limited space, we present the top 8 mentioned accounts in Table 3. The most highly mentioned account is JoVE the journal. Evernote’s Twitter account is ranked the second, revealing these Twitter users’ behavior of mentioning @MyEN to save the tweets of the article to their Evernote notebook stacks. In addition, two are university accounts, while four are researchers in disciplines related to JoVE’s publication scope.

Table 3. Top 8 most frequently mentioned Twitter users

<table>
<thead>
<tr>
<th>Mentioned User</th>
<th>No.</th>
<th>Coded Account Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>@JoVEJournal</td>
<td>366</td>
<td>JoVE the journal</td>
</tr>
<tr>
<td>@MyEN</td>
<td>57</td>
<td>Evernote (mention to save article)</td>
</tr>
<tr>
<td>@ColinW_Bell</td>
<td>18</td>
<td>research ecologist</td>
</tr>
<tr>
<td>@mwallenstein</td>
<td>18</td>
<td>soil microbial ecology professor</td>
</tr>
<tr>
<td>@crodarte</td>
<td>16</td>
<td>founder of a health-related company</td>
</tr>
<tr>
<td>@Penn</td>
<td>10</td>
<td>University of Pennsylvania</td>
</tr>
<tr>
<td>@Uutah</td>
<td>10</td>
<td>University of Utah</td>
</tr>
<tr>
<td>@MicahJMarty</td>
<td>9</td>
<td>coral reef ecologist</td>
</tr>
</tbody>
</table>
**Coding Schemes**

As is described in the Methods and Data section, we are still working on the content analysis. Due to limited space, we are only able to present our coding schemes in bullet points with the codes delimited by commas. The definition of the codes and more results will be presented in our later work.

Twitter accounts coding scheme:

- Account Gender: Male, Female, Not Sure, N/A
- Bot or NOT: Yes, No, Hybrid, Not Sure

Tweets’ text coding scheme:

- Elements Included: Whole Title, Part of Title, Summary of Title, Methodology of Article, Conclusion of Article, Concept in Article, Topic/Discipline of Article, Summary of Article, Comment on Article, Recommendation of Article, Author, Link Only.
- Sentiment: Positive, Negative.
- Self-promotion: Publisher, Individual, Institution, Colleague/Friend, Others.

**Discussion and Future Work**

In this study, instead of collecting tweets posted by a certain group of researchers or those sharing the same hashtags, we harvested our data by retrieving scientific article link citations. Therefore, these tweets “citing” articles from the same journal (JoVE) are not found to be as conversational or topical in terms of timeliness. However, they display their unique features. First of all, video/visual elements are highly mentioned in both hashtags and the tweets’ text. This is an advantage of video publications and could potentially facilitate the online scholarly communication of specific disciplines. In JoVE’s case specifically, the video articles are mostly video protocols that could potentially accelerate experimental biological, medical, chemical and physical research. Secondly, although the act of tweeting is a behavior to disseminate or discuss information, saving behavior (by mentioning @MyEN) has been observed. This demonstrates the possibility of capturing different types of digital traces that researchers leave on one single platform. Meanwhile, it also provides us more train of thoughts about studying how researchers’ information behaviors have been changed as the advent of new technology. Thirdly, bots play an important role in disseminating JoVE articles. However, does the large number of tweets generate impact or just empty buzz? Without further qualitative analysis, it is hard to know. As a result, we will pay specific attention to this aspect in the further content analysis. Last but not least, under the framework of altmetrics, whether and how much Twitter “citations” can reflect scholarly impact and societal impact are undergoing continuous discussion. When this study is done, more patterns of the tweets will be revealed, and the self-promotion behavior of different stakeholders will be better understood. This exploration will provide more evidence to the comprehension of altmetrics and impact.
Acknowledgments
The authors would like to thank Altmetric.com and its founder Euan Adie for supplying the data and its descriptions.

References
Normalization of zero-inflated data:
An empirical analysis of a new indicator family

Robin Haunschild\textsuperscript{1} Lutz Bornmann\textsuperscript{2}

\textsuperscript{1}r.haunschild@fkf.mpg.de
Max Planck Institute for Solid State Research, Heisenbergstr. 1, 70569 Stuttgart (Germany)

\textsuperscript{2}bornmann@gv.mpg.de
Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society,
Hofgartenstr. 8, 80539 Munich (Germany)

Abstract
Recently, two new indicators (Equalized Mean-based Normalized Proportion Cited, EMNPC, and Mean-based Normalized Proportion Cited, MNPC) were proposed which are intended for sparse data. We propose a third indicator (Mantel-Haenszel quotient, MHq) belonging to the same indicator family. The MHq is based on the MH analysis—an established method for polling the data from multiple 2×2 cross tables based on different subgroups. We test (using citations and assessments by peers, i.e., F1000Prime recommendations) if the three indicators can distinguish between different quality levels as defined on the basis of the assessments by peers (convergent validity). We find that the indicator MHq is able to distinguish between the quality levels in most cases while MNPC and EMNPC are not.

Conference Topic
Indicators

Introduction
Alternative metrics (altmetrics) have been established as a new fast-moving and dynamic area in scientometrics (Galloway, Pease, & Rauh, 2013). Initially, altmetrics have been proposed as an alternative to traditional bibliometric indicators. Altmetrics are a collection of multiple digital indicators which measure activity related to research papers on social media platforms, in mainstream media or in policy documents (National Information Standards Organization, 2016; Work, Haustein, Bowman, & Larivière, 2015). According to Haustein (2016), sources of altmetrics can be grouped into (i) social networks, (ii) social bookmarking and online reference management, (iii) social data (e.g., data sets, software, presentations), (iv) blogs, (v) microblogs, (vi) wikis, and (vii) recommendations, ratings, and reviews.

Recently, some indicators based on altmetrics have been proposed which are normalized with respect to the scientific field and publication year. These indicators were developed because studies have shown that altmetrics are—similar to bibliometric data—field- and time-dependent (see, e.g., Bornmann, 2014). Some fields are more relevant to the general public or a broader audience than other fields (Haustein, Larivière, Thelwall, Amyot, & Peters, 2014). The Mean Normalized Reader Score (MNRS) was introduced by Haunschild and Bornmann (2016) for normalization of data from social bookmarks and online reference management platforms (with a special emphasis on Mendeley readers), see also Fairclough and Thelwall (2015). The Mean Discipline Normalized Reader Score (MDNRS) was tailored specifically to Mendeley by Bornmann and Haunschild (2016b). The MDNRS uses Mendeley disciplines for field normalization. The employed normalization procedures rely on average value calculations across scientific fields and publication years as expectation values. Normalization procedures based on averages and percentiles of individual papers are problematic for zero-inflated data sets (Haunschild, Schier, & Bornmann, 2016). Bornmann and Haunschild (2016a) proposed the Twitter Percentile (TP)—a field- and time-normalized indicator for Twitter data.
The overview of Work, et al. (2015) on studies investigating the coverage of papers on social media platforms show that many platforms have coverages of less than 5% (e.g., Blogs, or Wikipedia). Erdt, Nagarajan, Sin, and Theng (2016) report similar findings in their meta-analysis. They found that former empirical studies dealing with the coverage of altmetrics show that about half of the platforms are at or below 5%; except for three (out of eleven) where the coverage is below 10%. Thus, altmetrics data are frequently concerned by zero-inflation.

Bornmann and Haunschild (2016a) circumvented the problem of zero-inflated Twitter data by including in the TP calculation only journals with at least 80% of the papers having at least 1 tweet each. However, this procedure leads to the exclusion of many journals from the TP procedure. Very recently, Thelwall (2017a, 2017b) proposed a new family of field- and time-normalized indicators for zero-inflated altmetrics data. The new indicators family is based on units of analysis (e.g., a researcher or institution) rather than on the paper level. They compare the proportion of mentioned papers (e.g., on Wikipedia) of a unit with the proportion of mentioned papers in the corresponding fields and publication years (the expected values). The family consists of the Equalized Mean-based Normalized Proportion Cited (EMNPC) and the Mean-based Normalized Proportion Cited (MNPC). Hitherto, this new family of indicators has only been studied on rather small samples.

In this study, we investigate the new indicator family empirically on a large scale (multiple complete publication years) and add a further indicator to this family. In statistics, the Mantel-Haenszel (MH) analysis is recommended for polling the data from multiple 2×2 cross tables based on different subgroups (here: mentioned and not-mentioned papers of a unit published in different subject categories and publication years compared with the corresponding reference sets). We call the new indicator Mantel-Haenszel quotient (MHq). In the empirical analysis, we compare the indicator scores with ratings by peers. We investigate whether the indicators are able to discriminate between different quality levels assigned by peers to publications. Thus, we test the convergent validity of the indicators. Since the convergent validity can only be tested by using citations (which are related to quality), the empirical part is based solely on citations and not altmetrics data. Good performance on the convergent validity test is a necessary condition for the use of the indicators in altmetrics (although for altmetrics, the relationship to quality is not clear).

**Indicators for zero-inflated count data**

Whereas the EMNPC and MNPC proposed by Thelwall (2017a, 2017b) are explained in the following two sections, the MHq is firstly introduced in the section thereafter. The next sections present not only the formulas for the calculation of the three metrics, but also the corresponding 95% confidence intervals (CIs). The CI is a range of possible indicator values: We can be 95% confident that the interval includes the “true” indicator value in the population. With the use of CIs, we assume that we analyse sample data and infer to a larger, inaccessible population (Williams & Bornmann, 2016). According to Claveau (2016) the general argument for using inferential statistics with scientometric data is “that these observations are realizations of an underlying data generating process … The goal is to learn properties of the data generating process. The set of observations to which we have access, although they are all the actual realizations of the process, do not constitute the set of all possible realizations. In consequence, we face the standard situation of having to infer from an accessible set of observations – what is normally called the sample – to a larger, inaccessible one – the population. Inferential statistics are thus pertinent” (p. 1233).

The relationship between CIs and statistical significance (in case of independent proportions) is as follows:
“1. If the 95% CIs on two independent proportions just touch end-to-end, overlap is zero and the p value for testing the null hypothesis of no difference is approximately .01.
2. If there’s a gap between the CIs, meaning no overlap, then p<.01.
3. Moderate overlap … of the two CIs implies that p is approximately .05. Less overlap means p<.05.

Moderate overlap is overlap of about half the average length of the overlapping arms” (Cumming, 2012, p. 402).

**Equalized Mean-based Normalized Proportion Cited (EMNPC)**

Thelwall (2017a, 2017b) introduced the EMNPC as an alternative indicator for zero-inflated count data. The approach of the EMNPC is to calculate the proportion of papers that are mentioned: suppose that publication set $g$ has $n_{gf}$ papers in the publication year and subject category combination $f$. $s_{gf}$ of the papers are mentioned (e.g., on Wikipedia). $F$ is defined as all publication year and subject category combinations of the papers in the set. The overall proportion of $g$’s papers that are mentioned is the number of mentioned papers ($s_{gf}$) divided by the total number of papers ($n_{gf}$):

$$p_g = \frac{\sum_{f \in F} s_{gf}}{\sum_{f \in F} n_{gf}} \quad (1)$$

However, $p_g$ could lead to misleading results if the publication set $g$ includes many papers which are published in fields with many mentioned papers. Thelwall (2017a, 2017b) proposes to avoid the problem by artificially treating $g$ as having the same number of papers in each publication year and subject category combination. The author fixes it to the arithmetic average of numbers in each combination, but recommends not including in the analysis combinations of $g$ with only a few papers. Thus, the equalized sample proportion of $g$, $\hat{p}_g$ is the simple average of the proportions in each combination

$$\hat{p}_g = \frac{\sum_{f \in F} s_{gf}}{\frac{n_{gf}}{|F|}} \quad (2)$$

The corresponding world sample proportion is defined as:

$$\hat{p}_w = \frac{\sum_{f \in F} s_{wf}}{\frac{n_{wf}}{|F|}} \quad (3)$$

In Eqns. (2) and (3), $[F]$ is the number of subject category and publication year combinations in which the group (in case of Eq. (2)) and the world (in case of Eq. (3)) publishes. Thus, the equalized group sample proportion has the undesirable property that it treats $g$ as if the average mentions of its papers do not vary between the subject categories. The EMNPC for each publication set $g$ is the ratio of both equalized sample proportions:

$$\text{EMNPC} = \frac{\hat{p}_g}{\hat{p}_w} \quad (4)$$

CIs for the EMNPC can be calculated as follows (Thelwall, 2017a):

$$EMNPC_L = \exp \left( \ln \left( \frac{\hat{p}_g}{\hat{p}_w} \right) - 1.96 \sqrt{\frac{(n_g - \hat{p}_g n_g) / (\hat{p}_g n_g)}{n_g} + \frac{(n_w - \hat{p}_w n_w) / (\hat{p}_w n_w)}{n_w}} \right) \quad (5)$$
Here, \( n_g \) is the total sample size of the group and \( n_w \) is the total sample size of the world.

**Mean-based Normalized Proportion Cited (MNPC)**

The second indicator proposed by Thelwall (2017a) has been named as Mean-based Normalized Proportion Cited (MNPC). The MNPC is calculated as follows: For each paper with at least one mention (e.g., on Wikipedia), the number of mentions is replaced by the reciprocal of the world proportion mentioned for the corresponding subject category and publication year. All other papers with zero mentions remain at zero. Let \( p_{gf} = s_{gf}/n_{gf} \) be the proportion of papers mentioned for publication set \( g \) in the corresponding subject category and publication year combination \( f \) and let \( p_{wf} = s_{wf}/n_{wf} \) be the proportion of world’s papers cited in the same year and subject category combination \( f \). Then:

\[
 r_i = \begin{cases} 
 0, & \text{if } c_i = 0 \\
 1/p_{wf} & \text{if } c_i > 0, \text{where paper } i \text{ is from year and subject category combination } f 
\end{cases}
\]

Following the calculation of the MNCS (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011), the MNPC is defined as

\[
 MNPC = \frac{(r_1 + r_2 + \cdots + r_n)}{n} 
\]

An approximate CI has been constructed by Thelwall (2016, 2017a) for the MNPC. In the first step, the lower limit \( L \) (MNPC\(_{gL}\)) and upper limit \( U \) (MNPC\(_{gU}\)) for group \( g \) in subject category and publication year combination \( f \) is calculated with:

\[
 MNPC_{gfL} = \exp\left(\ln\left(\frac{p_{gf}}{p_{wf}}\right) - 1.96 \sqrt{\frac{(n_{gf} - p_{gf}n_{gf})/(p_{gf}n_{gf})}{n_{gf}} + \frac{(n_{wf} - p_{wf}n_{wf})/(p_{wf}n_{wf})}{n_{wf}}} \right) 
\]

\[
 MNPC_{gfU} = \exp\left(\ln\left(\frac{p_{gf}}{p_{wf}}\right) + 1.96 \sqrt{\frac{(n_{gf} - p_{gf}n_{gf})/(p_{gf}n_{gf})}{n_{gf}} + \frac{(n_{wf} - p_{wf}n_{wf})/(p_{wf}n_{wf})}{n_{wf}}} \right) 
\]

In the second step, the group-specific lower and upper limits are used to calculate the MNPC CIs:

\[
 MNPC_L = MNPC - \sum_{f \in F} \frac{n_{gf}}{n_g} \left( \frac{p_{gf}}{p_{wf}} - MNPC_{gfL} \right) 
\]

\[
 MNPC_U = MNPC + \sum_{f \in F} \frac{n_{gf}}{n_g} \left( MNPC_{gfU} - \frac{p_{gf}}{p_{wf}} \right) 
\]

The MNPC cannot be calculated, if any of the world proportions are equal to zero. Furthermore, CIs cannot be calculated if any of the group proportions are equal to zero. Thus, Thelwall (2017a) proposed to remove the corresponding subject category publication year combination from the data or to add a continuity correction of 0.5 to the number of mentioned and not mentioned papers in these cases. We prefer the latter (to add 0.5 to the number of papers mentioned and not mentioned, respectively). This approach is recommended by Plackett (1974) for the calculation of odds ratios.
**Mantel-Haenszel quotient (MHq)**

For polling the data from multiple 2×2 cross tables based on different subgroups (which are part of a larger population), the most commonly used and recommended method is the MH analysis (Hollander & Wolfe, 1999; Mantel & Haenszel, 1959; Sheskin, 2007). According to Fleiss, Levin, and Paik (2003) the method “permits one to estimate the assumed common odds ratio and to test whether the overall degree of association is significant. Curiously, it is not the odds ratio itself but another measure of association that directly underlies the test for overall association … The fact that the methods use simple, closed-form formulas has much to recommend it” (p. 250). Radhakrishna (1965) demonstrate that the MH approach is formally and empirically valid against the background of clinical trials.

The MH analysis results in a summary odds ratio for multiple 2×2 cross tables which we call MHq. For the impact comparison of units in science with reference sets (the world), the 2×2 cross tables (which are polled) consist of the number of papers mentioned and not mentioned in subject category and publication year combinations \( f \). Thus, in the 2×2 subject- and year-specific cross table with the cells \( a_f, b_f, c_f, \) and \( d_f \) (see Table 1), \( a_f \) is the number of mentioned papers in subject category and publication year \( f \), \( b_f \) is the number of not mentioned papers in subject category and publication year \( f \), \( c_f \) is the number of mentioned papers published by group \( g \) in subject category and publication year \( f \), \( d_f \) is the number of not-mentioned papers published by group \( g \) in subject category and publication year \( f \). Note that the papers of group \( g \) are part of the papers in the world.

**Table 1. 2×2 subject-specific cross table**

<table>
<thead>
<tr>
<th>Group g</th>
<th>Number of mentioned papers</th>
<th>Number of not mentioned papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>World</td>
<td>( a_f )</td>
<td>( b_f )</td>
</tr>
<tr>
<td></td>
<td>( c_f )</td>
<td>( d_f )</td>
</tr>
</tbody>
</table>

We start by defining some dummy variables for the MH analysis:

\[
R_f = \frac{a_f d_f}{n_f} \quad \text{and} \quad R = \sum_{f=1}^{F} R_f, 
\]

(13)

\[
S_f = \frac{b_f c_f}{n_f} \quad \text{and} \quad S = \sum_{f=1}^{F} S_f, 
\]

(14)

\[
P_f = \frac{a_f + d_f}{n_f} \quad \text{and} \quad Q_f = 1 - P_f 
\]

(15)

Where: \( n_f = a_f + b_f + c_f + d_f \)

The MHq is simply:

\[
\text{MHq} = \frac{R}{S} 
\]

(16)

The CIs for MHq are calculated following Fleiss, et al. (2003). The variance of \( \ln \text{MHq} \) is estimated by:

\[
\tilde{V} \text{ar}[\ln(MHq)] = \frac{1}{2} \left\{ \frac{\sum_{f=1}^{F} P_f R_f}{R^2} + \frac{\sum_{f=1}^{F} (P_f S_f + Q_f R_f)}{RS} + \frac{\sum_{f=1}^{F} Q_f S_f}{S^2} \right\} 
\]

(17)

The confidence interval for the MHq can be constructed with

\[
\text{MHq}_L = \exp \left[ \ln(MHq) - 1.96\sqrt{\tilde{V} \text{ar}[\ln(MHq)]} \right] 
\]

(18)
\[ MH_{q_0} = \exp \left[ \ln(MHq) + 1.96\sqrt{Var[\ln(MHq)]} \right] \]  

(19)

Similar to the EMNPC and MNPC, it is an advantage of the MHq that the world average has a value of 1. It is a further advantage of the MHq that the result can be expressed as a percentage which is relative to the world average, e.g.: \( MHq = 1.30 \) means that the paper set under study has achieved an impact 30% above average.

Data sets used

Data sets used

We used the papers of the Web of Science (WoS) from our in-house database – derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), and Arts and Humanities Citation Index (AHCI) provided by Clarivate Analytics (formerly the IP and Science business of Thomson Reuters). All papers of the document type “article” with DOI published between 2010 and 2013 were included to study the indicators. Citations with a three-year citation window are retrieved from our in-house database (Glänzel & Schoepflin, 1995). For field classification, we used the overlapping WoS subject categories (Rons, 2012, 2014). In order to avoid statistical and numerical problems, we include only fields in the analysis where (1) at least 10 papers are assigned to and (2) the number of cited and uncited papers is non-zero. In total, these restrictions lead to a dataset including 4,490,998 papers.

We matched the publication data with peers’ recommendations from F1000Prime. F1000Prime is a post-publication peer review system of papers from mainly medical and biological journals. Papers are selected by a peer-nominated global “Faculty” of leading scientists and clinicians who then rate the papers and explain their importance. Thus, only a restricted set of papers from the papers in these disciplines covered is reviewed, and most of the papers are actually not. The Faculty nowadays numbers more than 5,000 experts worldwide. Faculty members can choose and evaluate any paper that interests them. The papers are rated by the Faculty members as “Recommended,” “Must read”, or “Exceptional” which is equivalent to recommendation scores (RSs) of 1, 2, or 3, respectively. Papers can be recommended multiple times. Therefore, we calculated an average RS (\( \bar{FRa} \)):

\[ \bar{FRa} = \frac{1}{i_{\text{max}}} \sum_{i=1}^{i_{\text{max}}} RS_i \]  

(20)

The papers are categorized depending on their \( \bar{FRa} \) value:

- Not recommended papers (Q0): \( \bar{FRa} = 0 \)
- Recommended papers with a rather low average score (Q1): \( 0 < \bar{FRa} \leq 1.0 \)
- Recommended papers with a rather high average score (Q2): \( \bar{FRa} > 1.0 \)

We only included fields where a paper with an F1000Prime recommendation is assigned to, following Waltman and Costas (2014). This reduces the total paper set included in the analysis to 2,873,476 papers.

Empirical analysis

The comparison of indicators with peer evaluation has been widely acknowledged as a way of investigating the convergent validity of metrics (Garfield, 1979; Kreiman & Maunsell, 2011). Convergent validity is the degree to which two measurements of constructs (here: two proxies of scientific quality), which should be theoretically related, are empirically related. Thelwall (2017b) justifies this approach as follows: “if indicators tend to give scores that agree to a large extent with human judgements then it would be reasonable to replace human judgements with them when a decision is not important enough to justify the time necessary for experts to read the articles in question. Indicators can be useful when the value of an assessment is not
great enough to justify the time needed by experts to make human judgements” (p. 4). Several publications investigating the relationship between citations and Research Excellence Framework (REF) outcomes report considerable relationships in several subjects, such as biological science, psychology, and clinical sciences (Butler & McAllister, 2011; Mahdi, d’Este, & Neely, 2008; McKay, 2012; Smith & Eysenck, 2002; Wouters et al., 2015). Similar results were found for the Italian research assessment exercise: “The correlation strength between peer assessment and bibliometric indicators is statistically significant, although not perfect. Moreover, the strength of the association varies across disciplines, and it also depends on the discipline internal coverage of the used bibliometric database” (Franceschet & Costantini, 2011, p. 284). The overview of Bornmann (2011) shows further results on journal peer review, that a higher citation impact of papers is to be expected with better recommendations from peers.

In recent years, the correlation between the F1000Prime RSs and citation impact scores has already been targeted. The results of the regression model of Bornmann (2015) demonstrate that about 40% of publications with RS=1 belong to the 10% most frequently cited papers, compared with about 60% of publications with RS=2 and about 73% of publications with RS=3. Waltman and Costas (2014) found “a clear correlation between F1000 recommendations and citations” (p. 433). The previous results on F1000Prime allow the prognosis, therefore, that citation-based indicators differentiate more or less clearly between the three RSs. In other words, the validity of new indicators can be questioned if the ability to differentiate is not given.

Against this backdrop, we investigate in the current study the ability of the three indicators for zero-inflated count data to differentiate between the RS groups. We start with the newly introduced MHq indicator. Figure 1 shows the MHqs with CIs for the three groups (Q0, Q1, and Q2) across four publication years.

Figure 1. MHqs with CIs for the three groups and four publication years. The horizontal line with MHq=1 is the worldwide average.

It is clearly visible that the MHq values are very different for the three groups which speak for their convergent validity: The mean MHq across the years is close to (but below) 1 for Q0. The mean MHq for Q1 is about eight times and that for Q2 is about 15 times higher than the...
mean MHq for Q0. It seems that the MHq indicator significantly separates between the different quality levels.
However, let us take a closer look at the MHq differences between the groups on the basis of their CIs following the rules of Cumming (2012) and Cumming and Finch (2005). If there is a gap between two CIs in the figure, then the difference is statistically significant ($p<.01$). This is the case for the years 2012 and 2013. Here, the indicator differentiates clearly and statistically significantly between the three groups ($p<.01$). In 2010 and 2011, there is also a statistically significant difference between Q0 and the other two groups. However, the CIs for Q1 and Q2 overlap in 2010 and 2011. If the overlap between the CIs is less than 50%, then the difference is statistically significant on the $p<.05$ level. This rule is reasonably accurate, however, when the two margins of error (length of one arm of a CI) do not differ by more than a factor of 2. The calculation of the overlaps yields an overlap of 43% in 2010 and 57% in 2011. Thus, the difference between the MHqs is statistically not significant in 2011 ($p>.05$). Although the difference is statistically significant in 2010 ($p<.05$), we cannot assume that the rule works accurately, because the two margins of error differ by a factor of 2.1.
The reason for the better result of the MHq in 2012 and 2013 than in 2010 and 2011 might be that 2012 and 2013 contain more uncited papers than 2010 and 2011. As MHq is designed for zero-inflated count data, a better performance can be expected for 2012 and 2013.

Figure 2. MNPC with CIs for the three groups and four publications years. The horizontal line with MNPC=1 is the worldwide average.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure2.png}
\caption{MNPC with CIs for the three groups and four publications years. The horizontal line with MNPC=1 is the worldwide average.}
\end{figure}
Figure 3. EMNPC with CIs for the three groups and four publications years. The horizontal line with EMNPC=1 is the worldwide average.

Figure 2 and Figure 3 show the results for the three groups for MNPC and EMNPC – the two indicators proposed by Thelwall (2017a). For both indicators, it is striking that all values in the graphs are very close to 1 – independent of the group. This is very different to Figure 1, in which the MHq values significantly differ from 1 for the two groups with recommendations (Q1 and Q2). This can be interpreted as a first sign that the MNPC and EMNPC do not differentiate between the quality levels in terms of FFa values.

The CIs in Figure 2 further reveal that the differences between the MNPCs for the different RS values are statistically not significant. There are clear and substantial overlaps for all CIs. The results in Figure 3 are very heterogeneous. In 2010, the mean value of Q2 is lower than the value of Q1. In 2013, the situation is reversed and in the expected direction then. In 2011 and 2012, the mean values are also in the expected direction, but there is a substantial overlap of the CIs (52% in 2012). According to the rules of Cumming (2012) and Cumming and Finch (2005), the differences between the CIs in in both years are statistically not significant.

Discussion

Although the empirical analyses in this study are based on citation data, the objective of the study is on developing indicators for sparse altmetrics data, i.e., zero-inflated altmetrics data. According to Neylon (2014), much of the data we have in altmetrics is sparse. An indicator with many zero values is unlikely to be informative about a scientific unit (e.g., a researcher or institution) in the first place (Thelwall, Kousha, Dinsmore, & Dolby, 2016). Thus, Thelwall (2017a, 2017b) proposed a new family of (meaningful) field- and time normalized indicators which are especially designed for the use with sparse data. The family consists of the EMNPC and MNPC indicators. Basically, the indicators compare the proportion of mentioned papers of a unit with the proportion of mentioned papers in the corresponding fields and publication years (the expected values).

The indicators of the family differ from most of the other indicators which have been proposed in bibliometrics and altmetrics hitherto. The other indicators are calculated for single publications and the user of the indicators can aggregate the indicator values (by averaging, summing, etc.). The indicators of the new family are not calculated for single
publications, but publication sets of groups (e.g., single researchers or institutes). Thus, these indicators cannot be used as flexible as the other bibliometric and altmetric indicators. However, we think that it will never be possible to develop reliable indicators with values for single publications for zero-inflated count data.

In this study, we analyzed the new indicator family empirically and added a further indicator variant – the MHq. We did not include altmetrics data in the empirical part of the study, although the indicator family focusses on them. Before the indicators can be used with altmetrics data, they have to be validated and this can only be done on the basis of citation data. Citation data allows formulating predictions which can be empirically validated with the new indicators. In this study, we tested with citation data whether the indicators are able to differentiate validly between three quality levels – as defined by F1000 RSs (FFa). Thus, we compared the indicator values with ratings by peers: Are the indicators able to discriminate between different quality levels assigned by peers to publications?

For the study, citations with a three-year citation window are retrieved from our in-house database as a compromise between having a significant correlation with quality (in the sense of post-publication peer assessments) and having a data set with rather many not mentioned papers. The results for the EMNPC and MNPC show that they cannot discriminate validly between the different quality levels. The scores for all quality levels are close to 1 (the worldwide average) and the CIs substantially overlap in many comparisons. Thus, the results point out that the convergent validity of the EMNPC and MNPC is not given. In this study, we further introduced the MHq to the new indicator family which is based on the established MH analysis. Since the MHq was able to discriminate empirically between the different quality levels – in most of the cases statistically significant – the convergent validity of the new variant seems to be given.

This study follows the important initiative of Thelwall (2017a, 2017b) to design new indicators for sparse data. Our study was the first independent attempt to investigate this indicator family empirically. Since this family is important especially for altmetrics data, we need further empirical studies which focus, e.g., on more sparse data than we used. Future empirical studies should investigate the new indicator family in other disciplines than biomedicine. F1000 focuses on the biomedical literature only.

Acknowledgements
The F1000Prime recommendations were taken from a data set retrieved from F1000 in November, 2017. We would like to thank Mike Thelwall for helpful correspondence regarding calculation of the CIs for MNPC and EMNPC.

References


M-Score: An Indicator Quantifying Individual's Scientific Research Output

Fei Shu

Abstract
This study introduces a new bibliometric indicator, m-score, quantifying individual’s scientific research output. m-score is based on the initiative of h-index using a single number to measure research performance but overcome its inconsistency in measurement. Comparing with the h-index, the m-score can measure research performance more consistently and accurately with the combination of the productivity and the quality.

Introduction
The h-index (Hirsch, 2005) was introduced by American physicist Jorge E. Hirsch in 2005. The index h is defined as “the number of papers with citation number ≥ h, as a useful index to characterize the scientific output of a researcher” (Hirsch, 2005, p. 16569). It is undoubtedly popular to use a single number to represent both the productivity and the quality of research (Bornmann, 2014) so that the h-index has been a popular research front in bibliometrics and scientometrics since its birth. As Figure 1 shows, Hirsch’s paper has been highly cited in the past 10 years; and a lot of variants of the h-index have also been developed (Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2010; Egghe, 2006; Jin, Liang, Rousseau, & Egghe, 2007; Lando & Bertoli-Barsotti, 2014; Ruane & Tol, 2008; Todeschini, 2011; Wan & Liu, 2014; Yan, Ding, & Sugimoto, 2011). On the other hand, it has been questioned by bibliometricians whether this single number could measure the research performance consistently and accurately (Bornmann, 2014; Bornmann, Mutz, & Daniel, 2008; Waltman & van Eck, 2012). The interest in the h-Index as well as its other derivatives has been declining since 2010 in Library and Information Science (LIS) as the proportion of citing papers from LIS literature has been decreasing as also shown in Figure 1.
Research Problem

Although Hirsch declares that the h-index could measure the impact of an individual’s research performance combining the productivity with the quality, and avoid the disadvantages of some existing indicators (e.g. total number of citations is too sensitive to few highly cited papers), the h-index does not take account of the disciplinary bias in terms of the citation rate and the influence of the academic age (Bornmann, 2014). In addition, it has been proved that the h-index might produce inconsistent ranking due to its arbitrariness (Waltman & van Eck, 2012).

As Waltman and van Eck (2012) argue, the h-index could be defined as “a scientist has an h-index of h if h of his publications each have at least 2h citations and his remaining publications each have fewer than 2(h+1) citations” or “A scientist has an h-index of h if h of his publications each have at least h/2 citations and his remaining publications each have fewer than (h+1)/2 citations” (p. 408), but these alternative definitions may lead to inconsistent results when calculating the h. As Figure 2 shows, scholar A has more publications (P) and receives more citations (C) compared with scholar B, but A’s h-index (11) is lower than B’s (12). Meanwhile, a contrary result may be returned when using either alternative definition mentioned by Waltman and van Eck as indicated in Figure 2. Indeed, every definition regarding calculating the h is arbitrary because the number of publications and the number of citations are two independent entities.

The purpose of this study is to try to develop a new indicator that use a single number use a single number to represent both the productivity and the quality of research. This new indicator could address all inconsistent issues of h-index and provide a more consistent and accurate measurement.

Figure 2. The inconsistency of the h-index in an example
**m-score**

For consistently and accurately measuring the research output, I propose a new indicator, m-score (m), defined as the maximum value of $m = \log_{10}(P) \times \log_{10}(C)$ where the number of publication (P) each has at least C citations.

As Figure 3 shows, for each scholar, we can present her/his research output in a citation-publication curve; every point on the curve represents that the scholar has P publications that were cited at least C times. In each given point B on the curve, the rectangle CBPO represents her/his research impact when the scholar has P publications that were cited at least C times. Based on the rectangle CBPO, we also can present a triangle 2C-2P-O that is exactly twice as big as the rectangle CBPO. The triangle 2C-2P-O can cover most of the research impact except for the two tails of the citation-publication curve that represent those highly cited papers and lowly cited or uncited papers, which may inflate the measurement of research output when simply counting the number of citations or publications (Hirsch, 2005). Since the m-score is calculated based on the area of the rectangle CBPO, it can avoid this inflation by excluding the both tails of the citation-publication curve.

**Figure 3. m-score and citation-publication curve**

In addition, comparing to the h-index arbitrarily requiring that P should be equal to C in calculation, m-score tries to reflect the maximum research impact with the combination of P and C as calculating the maximum value of the area of the rectangle CBPO. Even so one or

---

1 Please note that m-score is different from m-index, which is a variant of h-index where the lengths of academic careers was taken into account.
several extreme highly cited papers still could inflate the measurement. To avoid the inflation due to the outlier(s), we use the logarithms of both P and C instead of the absolute numbers when computing the m-score. All in all, the m-score measures the research impact based on a large group of frequently cited papers.

The computing of m-score is as follows: 1) rank the scholar’s publications based on the number of citations received in descending order; 2) assign the P value and C value representing the scholar has P publications that were cited at least C times; 3) produce the m as $m = \log_{10}(P) \times \log_{10}(C)$; 4) select the maximum value of m as the m-score. For example, 28 out of 32 papers published by Howard White (Drexel University) receive at least one citation. We ranked these publications in terms of the number of citations received and get the P value and C value representing the scholar has P publications that were cited at least C times as shown in Table 1. We also calculated the m as $m = \log_{10}(P) \times \log_{10}(C)$. The maximum value of m is 2,463 representing White has 16 papers that were cited at least 20 times.

Table 1 Computing of m-score for Howard White

<table>
<thead>
<tr>
<th>C value</th>
<th>P value</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>532</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>157</td>
<td>2</td>
<td>0.6610</td>
</tr>
<tr>
<td>123</td>
<td>3</td>
<td>0.9971</td>
</tr>
<tr>
<td>92</td>
<td>4</td>
<td>1.1823</td>
</tr>
<tr>
<td>87</td>
<td>5</td>
<td>1.3557</td>
</tr>
<tr>
<td>77</td>
<td>6</td>
<td>1.4680</td>
</tr>
<tr>
<td>71</td>
<td>7</td>
<td>1.5645</td>
</tr>
<tr>
<td>48</td>
<td>8</td>
<td>1.5183</td>
</tr>
<tr>
<td>30</td>
<td>9</td>
<td>1.4095</td>
</tr>
<tr>
<td>27</td>
<td>10</td>
<td>1.4314</td>
</tr>
<tr>
<td>26</td>
<td>11</td>
<td>1.4735</td>
</tr>
<tr>
<td>24</td>
<td>12</td>
<td>1.4895</td>
</tr>
<tr>
<td>23</td>
<td>13</td>
<td>1.5169</td>
</tr>
<tr>
<td>21</td>
<td>14</td>
<td>1.5154</td>
</tr>
<tr>
<td>21</td>
<td>15</td>
<td>1.5551</td>
</tr>
<tr>
<td>20</td>
<td>16</td>
<td>1.5666</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>1.4816</td>
</tr>
<tr>
<td>15</td>
<td>18</td>
<td>1.4763</td>
</tr>
<tr>
<td>13</td>
<td>19</td>
<td>1.4245</td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>1.3549</td>
</tr>
<tr>
<td>10</td>
<td>21</td>
<td>1.3222</td>
</tr>
<tr>
<td>7</td>
<td>22</td>
<td>1.1345</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>0.6497</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>0.4155</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>0.4208</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>0.4259</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>0.4309</td>
</tr>
<tr>
<td>1</td>
<td>28</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
Discussion and Conclusion

A test comparing the m-score with the h-index was conducted. Ten top LIS scholars were selected from the list provided by Cronin and Meho (2006) using the h-index to rank influential information Scientists. The number of their publications, the total citations they received, and their h-indexes were retrieved from the Scopus database. As Table 2 shows, although the rank by m-score is similar to the rank by h-index, the m-score could differentiate their research performance better than the h-index while there are three tie in the h-index of 23 and two tie in the h-index of 19. Also, the m-score favours scholars with a large set of frequently cited papers rather than a small number of highly cited papers. For example, McCain has more publications and receives more citations than Fidel but her m-score is lower than Fidel. McCain’s advantage in citation comes from a highly cited paper receiving 532 citations while Fidel’s most highly cited paper were cited 225 times. When the influence of the highly cited papers was excluded, Fidel has 11 papers receiving at least 46 citations while McCain only has 10 papers that were cited at least 33 times. Thus, Fidel’s m-score is higher than McCain’s.

In summary, this study proposes a new indicator, m-score, based on Hirsch’s initiative using a single number to measure research output. As the preliminary test indicate, compared with the h-index, the m-score could measure the research output more consistently and accurately with the combination of the productivity and the quality. In the future, the m-score should be tested in large and various data sets for its validity and reliability.

Table 2. Comparison between h-index and m-score in Top 10 LIS scholars

<table>
<thead>
<tr>
<th>Name</th>
<th># of publications</th>
<th># of citations received</th>
<th>h-index</th>
<th>Rank by h-index</th>
<th>m-score</th>
<th>Rank by m-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amanda Spink</td>
<td>199</td>
<td>6,358</td>
<td>40</td>
<td>1</td>
<td>2.5827</td>
<td>1</td>
</tr>
<tr>
<td>Nicholas Belkin</td>
<td>104</td>
<td>2,921</td>
<td>27</td>
<td>2</td>
<td>2.3252</td>
<td>2</td>
</tr>
<tr>
<td>Tefko Saracevic</td>
<td>78</td>
<td>3,886</td>
<td>23</td>
<td>3</td>
<td>2.1660</td>
<td>3</td>
</tr>
<tr>
<td>Blaise Cronin</td>
<td>168</td>
<td>2,047</td>
<td>23</td>
<td>3</td>
<td>1.9388</td>
<td>4</td>
</tr>
<tr>
<td>Gary Marchionini</td>
<td>122</td>
<td>2,288</td>
<td>23</td>
<td>3</td>
<td>1.9218</td>
<td>5</td>
</tr>
<tr>
<td>Christine Borgman</td>
<td>88</td>
<td>1,923</td>
<td>19</td>
<td>6</td>
<td>1.8335</td>
<td>6</td>
</tr>
<tr>
<td>Marcia Bates</td>
<td>45</td>
<td>2,218</td>
<td>19</td>
<td>6</td>
<td>1.8124</td>
<td>7</td>
</tr>
<tr>
<td>Raya Fidel</td>
<td>37</td>
<td>1,064</td>
<td>14</td>
<td>10</td>
<td>1.7316</td>
<td>8</td>
</tr>
<tr>
<td>Howard White</td>
<td>32</td>
<td>2,463</td>
<td>16</td>
<td>9</td>
<td>1.5666</td>
<td>9</td>
</tr>
<tr>
<td>Katherine McCain</td>
<td>64</td>
<td>1,311</td>
<td>17</td>
<td>8</td>
<td>1.5185</td>
<td>10</td>
</tr>
</tbody>
</table>

Acknowledgments

This study is supported by iFellows Doctoral Scholarship provided by the Andrew W. Mellon Foundation and the Doctoral Research Scholarship provided by the Fonds de recherche société et culture Québec (FRQSC).

References


Bornmann, L., Mutz, R., & Daniel, H. D. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using


Lando, T., & Bertoli-Barsotti, L. (2014). A New Bibliometric Index Based on the Shape of the Citation Distribution. *PLoS ONE, 9*(12). doi:10.1371/journal.pone.0115962


Discrimination Measurement Method
on H-index and G-index Using Jain’s Fairness Index

Adian Fatchur Rochim¹, Abdul Muis² and Riri Fitri Sari³

¹adian.fatchur@ui.ac.id, ²muis@ui.ac.id, ³riri@ui.ac.id

Department of Electrical Engineering, Faculty of Engineering
Universitas Indonesia, Depok, 16424, Indonesia

Abstract
In this paper observes discrimination of the well-known indices, H-index and G-index. There are many
definitions of impact of the researchers in the literature, such as citations, patents, books, source codes of the
open source, and social media. H-index is an alternative tool to measure performance of the researchers. Because
most of the database indexers use H-index to provide the figure of the researcher performance, so that H-index
many used for comparing the study progress among students or for considering research funding. On the other
hand, it was found that H-index has some weaknesses for calculating the researchers who have a few papers with
the high number of citations, and those with more papers but few citations. Many proposals suggested index
variants to improve the H-index calculation method. No other indices substitute the H-index, because the
 correlation of the other indices to the H-index is still relatively high. In this work, we propose a
discrimination tool to show the discriminate of the H-index. The discrimination tool method based on Jain’s
fairness index. Two indexing methods H-index and G-index are used as samples to compare the discrimination.
To prove the effectiveness of the propose method uses the citation set data of 1,000 researchers gained from
Scopus by application programming interface (API). We find that the discrimination of the H-index has the
average of 23.16% and 14.66% for G-Index for H-index range 1-216. We find that the discriminator tool as an
alternative tool for accommodating the discrimination calculation of the indices.

Keywords: discrimination, indices, H-index, G-index.

Conference Topic
Indicators

1. Introduction
The use of the citation records of research results has grown rapidly. This condition is a
portion of the contribution of the database indexers such as Thomson Reuters Web of Science
(WoS), Scopus, Google Scholar, Microsoft Academic Graph, Directory of Open Access
Journals (DOAJ), Connecting Repositories (CORE) and others. The Indexing system is to
identify papers in journals, conference proceedings, and institutional repositories. Scopus,
WoS, and Google Scholar use H-index to measure the productivity of researchers. Hirsch in
2005 proposed the combination of paper number and citation number to calculate the index
value of the researcher widely known as H-index (Hirsch, 2005). Despite a lot of proposals
have been submitted for improving the H-index calculation method, at the 10th anniversary of
the H-index, this method it is still used to measure the performance of researchers (Bornmann,
2011). H-index is also used to measure the credibility of a journal (Schubert & Glanzel, 2007)
(Science & Leuven, 2006)(Torra & Narukawa, 2008). H-index is as well used to justify
research funds granting and compare students' progress (Kreiner, 2016).

H-index has been identified with some weaknesses. In 2011, Lutz Bornmann stated that those
beginners in research are having difficulty to get citation. On the other hand, for senior
researchers with a lot of publications will have their citation number continue to grow
although they are no longer producing paper (Bornmann, 2011). Radko Mesiar in 2016, stated
that H-index was not sensitive to the condition in which the researcher has the large number
of papers with a small number of citations namely big produce researchers (Gagolewski & Grzegorzelowski, 2009). This is also the case with researchers with a small number of papers with a huge number of citations namely perfectionist researchers (Mesiar & Gagolewski, 2016). Alonso in 2009 stated that H-index could not be used to compare a senior researcher with a new one. For beginner researchers, the addition of citations will not significantly raise their H-index (Alonso, Cabrerizob, Herrera-Viedmac, & F.Herrera, 2009). Egghe in 2006 stated that to measure the performance of a researcher, one could consider the overall number of citations obtained, especially for highly cited papers (Egghe, 2006). Quentin L. Burrell in 2013 used a standard size–frequency distribution to compare H-index and G-index. They also showed that the top cited paper is not significantly contributed to the H-index value (Burrell, 2013). Bornman, in 2011 compared the H-index and their varians by using meta-analysis study. He stated that the H-index, and its family have the mean correlation coefficient that varies between 0.8-0.9 (Bornmann, 2011).The H-index method is based on the number of citations and the number of papers. Another index is G-index method, which is based on the Lotka’s Law. Both of methods became calculation index useful tools to indicate the achievement of researchers' performance.

The definition of fairness in Cambridge Dictionary is “the quality of treating people equally or in a way that is right or reasonable” (Cambridge, 2017). In this work, the fairness of the indices method based on Jain’s fairness index method. The definition of fairness index was introduced by Raj Jain in 1984, which is a function to calculate the fairness value.

Objectives and Method
From literatures, we found that the H-index is insensitive to calculate authors who have a few papers with a high number of citations, and those with more papers but fewer citations. So we need to measure the fairness index for the H-index and variants method.

The aim of this work is to find parameters and a method to measure the discrimination index of H-index and variants, based on the fairness using Jain’s method. We performed some experiments using data, which is gained from Scopus for selected researchers' data. The list of the author name gained from the webometrics of the ranking of researchers page.

To provide the figure of the discrimination for this work, we used the calculation comparison sample of the H-index calculation value. The parameters to be compared are citation, geometric area, and the number of the papers.

For example, we have citation set data of two researchers A and B. The researcher A has higher index than the index of the researcher B, although the geometric area of the citation set of A is lower than the citation set of B. The discrimination index of B will a higher than the discrimination index of A, because the index result of B is lower than the index of A. Figure 2 shows the comparison of H-index from the citation set data. Data set is gained from the Scopus database 9 February 2017.

Figure 2. shows that Egghe has the H-index of 24 which is lower than that of Sangwal's H-index (26), although the top cited papers and the number of papers of Egghe more than Sangwal. So that is the term of the discrimination is used. Therefore, it is important to measure the level of discrimination of the indices' method. Finally, we can say that the discrimination index is the difference tolerance degree on the index calculation method. On the occasion where index value has a smaller discrimination index, meaning that is closer to
the fairly index value. In Figure 2, we can see the bias that the top cited paper of citations set data of Egghe is more than Sangwal, but the H-index of Sangwal, which is 26 is higher than index of Egghe (24). The first hypothesis is that the highest citation number of paper of an author does not correlate directly with the high value of H-index. The second hypothesis is that the number of indexed papers on the lower side of the H-index-line does not contribute to the increase of the H-index.

Figure 2. H-index Comparison of the citation set data of the Sangwal and Egghe publications.

Bornmann in 2011 stated that a paper with no citation cannot increase the H-index value (Bornmann, 2011). Formerly, Egghe in 2006 expressed that the top cited paper is important to calculate the researcher performance (Egghe, 2006). Furthermore the performance of an author should consider the area of the lower and higher side of the H-index-line. This work describes the important aspect of the discrimination measurement to provide the fairness for researchers performance calculation. Calculation method has been considered for calculating the top cited articles. We noted some problems in the calculation of an author index. Condition in which new papers are not cited and therefore have a low citation number. Afterwards, for improving the sensitivity of the indices on the next research, firstly, we require to design a measurement tool to calculate the level of discrimination.

The discrimination index value is obtained from some vectors such as the absolute and real vector obtained (Jain et al., 1984). It has a value between 0 and 1. The discrimination index is the opposite of the fairness index (Jain et al., 1984).

\[
\text{Discrimination Index} = 1 - \text{Fairness Index} \quad (1)
\]

The definition of fairness index here is perceived as a factor of fairness received and represented into a number of properties (Jain et al., 1984). Fairness index can be calculated as follows:

\[
f(x) = \frac{\sum_{i=1}^{n} x_i}{N \sum_{i=1}^{n} x_i^2} \quad \text{where} \quad x_i \geq 0 \quad (2)
\]
H-index Method
To calculate the H-index value, we use Choquet and Sugeno integral approach. The method explain as follows. The membership function expressed the degree of membership in a set. The value of this function is the interval \([0,1]\), and is expressed as \(\mu_A\).

\[\mu_A: x \rightarrow [0, 1]\] (3)

The membership function expressed as \(\mu_A(y)\) valued as 1, if \(x\) is a full member of the set \(A\). Meanwhile if the degree of membeship is in the interval of 0 and 1, for example \(\mu_A(y) = \mu\), and \(y\) expressed as the member of the set \(A\) with the membership degree of \(\mu\) (Zadeh, 1965). Fuzzy integrals are defined as a function at a reference value of \(X\), where \(X = \{y_1,y_2,...,y_n\}\), is a finite reference set of the papers (Torra & Narukawa, 2008).

Then the pair is set sequentially stating the fuzzy sets of:

\[A = \{(x_1, \mu_A(x_1)), (x_2, \mu_A(x_2)), ..., (x_n, \mu_A(x_n))\}\] (6)

Vicence Torra in 2008 defined that the h-index can be approximated by the Fuzzy equation the Sugeno and Choquet Integral approaches (Torra & Narukawa, 2008).

**H-index** is a function of \(X : S \rightarrow \mathbb{R}\), so that

\[X(y_1, ..., y_n) = \{\max\{h = 1, ..., n: y_h \geq h\} \text{ jikai } y_1 \geq 1, \} \text{ otherwise} \] (7)

The number of citation (JS)-index is defined as the number of citation of author \(p\). \(Y_p\) is a set of papers owned by \(p\) while \(f(y)\) is the set of citations of paper \(x \in Y_p\). Therefore \(JS_p\) is defined as (Torra & Narukawa, 2008):

\[JS_p = \sum_{x \in Y_p} f(x)\] (8)

a researcher has an h-index \(h\), if \(h\) is the number of papers cited as a number as \(h\) per paper. Using \(Y_p\) and \(f\) definition above, the h-index calculation of the researcher \(p\) can be written as:

\[h_p = \max_i \min_j (f(x_{\partial(i)}), j)\] (9)

where \(\{\partial(1), ..., \partial(N)\}\) is a permutation of \(\{1, ..., N\}\), therefore \(f(x_{\partial(1)}) \geq f(x_{\partial(2)}) \geq \cdots \geq f(x_{\partial(N)})\). Algorithm 1. shows the H-index calculation algorithm which is defined based on the Choquet and Sugeno integral approaches. The result calculation of the H-index method based on the methods approaches significantly same as the H-index result of the Scopus.

G-index Method
The g value on the G-index method is the highest number of papers with the number of citations. This is greater than the square of the paper numbers, that is ordered by the sequence number of the papers sorted from those with the largest to smallest citation (Burrell, 2009). G-index theory was introduced by Leo Egghe in 2006, to fix the h-index of Hirsch (2005). The measurements are taken from the G-index of citations is obtained, “g-index is the largest number of citations of articles \(g\) with papers that has been highly cited at least \(g^2\)”. It has been demonstrated that the value of \(g\) is unique for each set of articles, and therefore it is obtained that \(g \geq h\) [9]. G-index method has been applied to the data of the work of researchers, and the result comparison of the g of the G-index is correlated with the h of H-index (Bornmann, 2011)(Egghe, 2006). G-index takes into account the number of papers, which have the top citation.
Algorithm 1. The H-index calculation algorithm using the Choquet and Sugeno Integral approaches.

<table>
<thead>
<tr>
<th>Line</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>global variables</td>
</tr>
<tr>
<td>2</td>
<td>( Y_p ), number of paper</td>
</tr>
<tr>
<td>3</td>
<td>( JS ), number of citation per paper</td>
</tr>
<tr>
<td>4</td>
<td>( P ), paper number</td>
</tr>
<tr>
<td>5</td>
<td>( P_s ), array of paper number (( P, JS ))</td>
</tr>
<tr>
<td>6</td>
<td>( P1 ), min variable</td>
</tr>
<tr>
<td>7</td>
<td>( H-index ), h-index value</td>
</tr>
</tbody>
</table>

```plaintext
1:
for \( i = 1 \) to \( Y_p \) do
  swap \( P_s(j) \), \( P_s(j+1) \)
end for
1:
for \( i = 1 \) to \( Y_p \) do
  \( i = \min P_s(i) \)
end for
1:
for \( i = 1 \) to \( Y_p \) do
  \( H-index = \max P1(i) \)
end for
```

The general formula Lotka’s Law used in g-index is shown in equation (10):

\[
g = \left( \frac{\alpha-1}{\alpha-2} \right)^{\frac{\alpha-1}{\alpha}} \frac{1}{T^{\alpha}}
\]  

(10)

The calculation results of the algorithm was compare the result in the Egghe method (Egghe, 2006) that the result is same value.

Figure 3. shows the representation of the G-index calculation based on Lotka’s Law. Lotka’s Law in which \( \alpha > 2 \) is the exponent of Lotkaian and \( T \) is the number of data sources. G-index method has been applied to the data of the work of Nobel laureates, and the comparison result that the \( g \) of the G-index has the correlation with the \( h \) of H-index (Bornmann, 2011)(Egghe, 2006). G-index takes into account the number of papers, which have the top citation. The G-index has a strong correlation with the H-index. Therefore, G-index is used for comparing the H-index in order to make the measurement tools of fairness index to this work. Algorithm 2. shows the G-index calculation algorithm which is based on the Lotka’s Law.

Our discrimination index method can be explain as follows:

1. Getting list of the author name from the webometrics site.
2. Querying Scopus ID of the authors from Scopus database using application
programming interface (API).
3. Parsing the data.
4. Querying the authors’ paper id of each author id.
5. Querying the number of citation of each paper id of the authors.
6. Calculating the H-index and G-index.
7. Obtaining the default value of the author H-index area. For example an author within an index of 7. It means that the default value of the total number of papers cited is 7, and the citation number 7 for each papers.
8. Calculating the geometric area of each author data,
9. Calculating the total number of the paper cited.
10. Calculating the discrimination index of authors.
11. Representing the result.

Table 1. shows the comparison of the calculation the H-index and G-index from the citation set data of the author name Egghe (accessed 9 December 2016 from Scopus). H-index value of 21 and G-index value of 36 for Egghe citation set data. Jain’s method measure the fairness index of resources allocation. The resources to evaluate the researcher performance is the total number of paper citation, the total number of paper indexed and area of the author data in the H-index graph. Jain method is chosen because it has advantages that it can be implemented on unlimited data and have an independent scale as the unit of measurement, bounded value of 0 and 1, It also has the continuous characteristic that can be presented by the percentage. The method can detect the slightest changes (continuous) that is appropriate to be applied to measure the discrimination value.

Algorithm 2. The G-index calculation algorithm based on Lotka’s Law

```plaintext
1: global variables
2: Y_p, number of paper
3: JS, number of citation per paper
4: P, paper number
5: P_s, array of paper number (P, JS)
6: P_1, min variable
7: H-index, h-index value
8: end global variables
9: for i=1:Y_p do
10: for j=0:Y_p-1 do
11: if JS(P_s(j)) > JS(P_s(j+1)) then
12: swap P_s(j), P_s(j+1)
13: end if
14: end for
15: end for
16: for i=1:Y_p do
17: G1 = P_s(i) * P_s(i)
18: G2 = JS(i) + JS(i-1)
19: end for
20: for i=1:Y_p do
21: if G1(i) ≥ G2(i) then
22: end for
```

471
Figure 3. The number of citation vs square of paper number to represent G-index calculation by Lotka’s Law

Table 1. Calculation of the H-index and G-Index

<table>
<thead>
<tr>
<th>Paper Number</th>
<th>TC</th>
<th>r</th>
<th>ΣTC</th>
<th>r²</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>224</td>
<td>1</td>
<td>224</td>
<td>1</td>
</tr>
<tr>
<td>2.</td>
<td>122</td>
<td>2</td>
<td>346</td>
<td>4</td>
</tr>
<tr>
<td>3-18</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>22</td>
<td>19</td>
<td>1010</td>
<td>361</td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>21</td>
<td>1053</td>
<td>441</td>
</tr>
<tr>
<td>22</td>
<td>20</td>
<td>22</td>
<td>1073</td>
<td>484</td>
</tr>
<tr>
<td>23</td>
<td>20</td>
<td>23</td>
<td>1093</td>
<td>529</td>
</tr>
<tr>
<td>24-33</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>15</td>
<td>34</td>
<td>1279</td>
<td>1156</td>
</tr>
<tr>
<td>35</td>
<td>14</td>
<td>35</td>
<td>1293</td>
<td>1225</td>
</tr>
<tr>
<td>36</td>
<td>14</td>
<td>36</td>
<td>1307</td>
<td>1296</td>
</tr>
<tr>
<td>37</td>
<td>14</td>
<td>37</td>
<td>1321</td>
<td>1369</td>
</tr>
<tr>
<td>38</td>
<td>14</td>
<td>38</td>
<td>1335</td>
<td>1444</td>
</tr>
</tbody>
</table>

TC = Total Citation, the number of citation
r = the number of paper (ordered by highest to lowest citation)
ΣTC = Sum of TCᵢ + TCᵢ₋₁, r² = square of r
Range numbers of 3 – 18 and 25 – 33 are not written to shorten the table.

The source data of the citation set data gained from the Scopus databases of 1000 world top scientists of the webometrics (February 2017). Each query maximum 200 counts and default 25 maximum query. The quota of the maximum query 20,000 each 7 days on the Scopus Data. The data used in this work is 1000 world top scientists of the webometrics version, and 500 papers of each. We make a query program for querying by Python language. The index fairness can be defined as:

Discrimination index is provided by:

\[
discrimination = (1 - fairness \ index) \times 100\%
\]  

(11)

Resources allocation in this work are defined as resources as 1) the index value, 2) the number
of the cited papers, and 3) the Geometric area. The Geometric Area (GA) is the sum of the papers and citations multiplication. The areas identified as number as figure 1 are used as the base of our calculation. Fairness is an important performance criterion in all resource allocations (Jain, Chiu, & Hawe, 1984). Figure 1 shows the sample of the geometric area of a researcher. The GA is sum of the areas a,b,1,b,1,b,2,c,1,c,2,d,1,d,2,e,1,e,2,f,1,f,2,g, and h.

\[ GA = a + (b_1 + b_2) + (c_1 + c_2) + (d_1 + d_2) + (e_1 + e_2) + f_1 + f_2 + g + h \]  

(4)

Figure 1. The geometric area (GA) of the citation data set of the researcher.

The Geometric area of the researchers can be written as follows:

\[ GA = \sum_{i=1}^{n}(x_i + (0.5 * y_i)), \text{ where } y_i = x_{i-1} - x_i \]  

(5)

in which \( x_i \) = the number of citation of the paper-i

We proposed the parameters for calculating the discrimination index as follows:

1. \( \alpha_t \) is the geometric area (GA) of the author data. While the default of the GA of the index value is \( \alpha_\circ \).
   The \( \alpha_t \) value is the GA of the citation set data of the authors, and the \( \alpha_\circ \) is the value of the default GA or the square of the H-index. For example the \( \alpha_\circ \) H-index of 7 is 49 and. Each the default GA depend on the H-index value.

The Geometric are of the researchers can be written as follows:

\[ GA = \sum_{i=1}^{n}(x_i + (0.5 * y_i)), \text{ where } y_i = x_{i-1} - x_i \]  

(12)

\( x_i \) = the number of citation of the paper-i

2. \( \beta_t \) is the H-index or the other indices value of the authors and \( \beta_\circ \) is the index value of the H-index from the citation set data of the authors.

3. Total number of the paper cited is \( \gamma_t \) and the default number of paper cited is \( \gamma_\circ \). The default number of paper cited depends on the H-index value, for example \( \gamma_\circ \) =7, if the number cited of the paper of 7. Likewise \( \gamma_\circ \) = 10, for the number of the cited papers of 10.

Discrimination index can be calculated as:
\[ \text{Discrimination index} = 1 - \left( \frac{|\alpha + \beta + \gamma|^2}{3 + (\alpha^2 + \beta^2 + \gamma^2)} \right) \]  

\[ \alpha = \alpha_t / \alpha_o \]  

\[ \beta = \beta_t / \beta_o \]  

\[ \gamma = \gamma_t / \gamma_o \]  

For sample case of the discrimination index calculation of two researcher A and B explain as follows. For example, the citation set data of author A with 6 papers with citation set \((6,5,4,3,2,1)\), the citation set data of author B with 4 papers with citation set \((4,4,4,4)\). The authors has same H-index of 4.

The default citation set for each paper of H-index 4 is \((4,4,4,4)\).

1. Discrimination index calculation Author of A:

\[ \alpha_t = (6 + (5 + 0.5) + (4 + 0.5) + (3 + 0.5) + (2 + 0.5) + (1 + 0.5)) = 23.5 \]

\[ \alpha_o = 4^2 = 16 \]

\[ \beta_t = 4; \beta_o = 4 \]

\[ \gamma_t = 6; \gamma_o = 4 \]

\[ \alpha = \frac{23.5}{16} = 1.47 \]

\[ \beta = \frac{4}{4} = 4 \]

\[ \gamma = \frac{6}{4} = 1.5 \]

\[ \text{discrimination index} = (1 - \left( \frac{(1.47 + 4 + 1.5)^2}{3 + (1.47^2 + 4^2 + 1.5^2)} \right) = 1 - 0.79 = 0.21 = 21\% \]

Therfore discrimination Index for author A = 21\%.

2. Discrimination index calculation Author B:

\[ \alpha_t = (4 + 4 + 4 + 4) = 16 \]

\[ \alpha_o = 4^2 = 16 \]

\[ \beta_t = 4; \beta_o = 4 \]

\[ \gamma_t = 4; \gamma_o = 4 \]

\[ \alpha = \frac{16}{16} = 1 \]

\[ \beta = \frac{4}{4} = 1 \]

\[ \gamma = \frac{4}{4} = 1 \]

\[ \text{discrimination index} = (1 - \left( \frac{(1 + 1 + 1)^2}{3 + (1^2 + 1^2 + 1^2)} \right) = 0 = 0\% \]

Therefore discrimination Index for author B = 0\%, or absolutely fair.

**Result and Discussion**

We find that the discrimination index of the H-index ranges from 1-80 and 181-216 of H-indexes are higher than the range 81-161 of H-index. Table 1 shows that the discrimination index range 1-80 and 181-216 have 25-36\%. It means that the H-index method insensitive for the researcher with small number of papers with huge number of citations (Bornmann, 2011)(Torra & Narukawa, 2008)(Egghe, 2006).

It was found that the measurement of the discrimination related with higher value for the H-index range 1-80 and 182-216 than the H-index range 81-180. The results demonstrated that the lower and upper areas of the H-index for researcher usually has a fairly wide area.
However this is not contributing to the H-index. The mean index of discrimination for the H-index are 25% and 19% for the G-index. The H-index discrimination is higher than the G-index. The research result of the index value discrimination of G-index is lower than for H-index. The discrimination index of the G-index is lower than H-index, this finding correlated with the G-index that has considered those top cited papers of the authors.

Table 2. Fairness and discrimination index of the authors (accessed 30 March 2017)

<table>
<thead>
<tr>
<th>Range of H-index</th>
<th>Fairness</th>
<th>Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H-index</td>
<td>G-index</td>
</tr>
<tr>
<td>1-20</td>
<td>0.68</td>
<td>0.79</td>
</tr>
<tr>
<td>21-40</td>
<td>0.76</td>
<td>0.94</td>
</tr>
<tr>
<td>41-60</td>
<td>0.74</td>
<td>0.84</td>
</tr>
<tr>
<td>61-80</td>
<td>0.74</td>
<td>0.77</td>
</tr>
<tr>
<td>81-100</td>
<td>0.77</td>
<td>0.84</td>
</tr>
<tr>
<td>101-120</td>
<td>0.78</td>
<td>0.85</td>
</tr>
<tr>
<td>121-140</td>
<td>0.78</td>
<td>0.85</td>
</tr>
<tr>
<td>141-160</td>
<td>0.83</td>
<td>0.88</td>
</tr>
<tr>
<td>161-180</td>
<td>0.86</td>
<td>0.94</td>
</tr>
<tr>
<td>181-216</td>
<td>0.74</td>
<td>0.84</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4. shows that both indices have the same pattern that is relatively high on the range of the H-index 1-80, this mean the measurement method for the calculation index should be improved to get more sensitive for author that has more paper and low citation.
Conclusion
It was found that the discrimination occurs especially in terms of the citation set data of the big produce researchers and perfectionist researchers. The data processed involves three parameters, namely: the index value obtained, the geometric area, and the number of the cited papers. The method can be used to measure the level of fairness for the H-index and G-index. This approach of the discrimination index can be used to compare the insensitivity and discrimination of those well known indices. The mean values from the H-index discrimination calculation from the existing data are 23.16 % for H-index and 14.66 % for G-Index. Therefore the G-index and H-index should be improved to get a fairer result.

Acknowledgments
This research is funded by the grant from the Ministry of Research and Higher Education of the Republic of Indonesia number 3202/UN2.R3.1/PPM.00.01/2017.

References
Allocating the Credit in High Energy Physics Collaborative Research

Deng Siyi¹  Chen Chong ²  Zhang Jingying ³

¹siyideng@163.com
Beijing Normal University, Beijing (China)

²chenchong@bnu.edu.edu
Beijing Normal University, Beijing (China)

³201511260102@mail.bnu.edu.cn
Beijing Normal University, Beijing (China)

Abstract

Based on some heuristic methods, this study improved the credit allocation in three aspects. For one thing, combining the authors' self-citation with the co-cited publications; secondly, employing different allocation rules to different range of authors' number; lastly, comparing the research experience of authors with the content of the target article. The improved model has been compared with several methods by statistical analysis. The main conclusions are as follows: 1) considering the co-citation, the methods based on the signature order are greatly improved at a confidence level of 90%; 2) self-citation can make up for the lack of the co-citation; 3) the content matching has a significant impact on the credit allocation at a confidence level of 97%. Therefore, under the premise of reasonable results, the "co-citation + self-citation + research content matching” model can effectively improve the credit allocation method.

Conference Topic

The application of informetrics on evaluation; co-authorship credit allocation; content matching; co-citation.; self-citation.

Introduction

Allocating credit among co-authors of a publication is a basic issue to many scientometric studies, such as the academic impact of researchers, and the capacity evaluation of institutes or countries. However, there are some problems that cannot be overlooked, which making the research challenging:

(1) Differences in authors’ signature rules. In some disiplines, the authors who sign in the first place tend to receive more attention(Van Praag M & Van Praag B M S,2007), and the order of the authors' signature usually determines how much the author earns, and in other disiplines, the authors' signature is only in alphabetical order, such as high-energy physics(Blaise Cronin,2001). Besides, even in the same displine that may be different (Joseph K & Laband D N & Patil V, 2005). Therefore, although there are many credit allocation algorithms developed based on the signature order, no one can cover all the rules.

(2) "Hyperauthorship" phenomenon. In recent years, the number of articles more than 50 authors has been significantly increased, especially greater than 1000, which is common in the field of biomedical and high-energy physics. Blaise Cronin (2001) has proposed a new concept on this phenomenon - "hyperauthorship", which means massive authorship levels. The emergence of this new phenomenon has changed authors’ cooperation mode (Birnholtz J P, 2006): the contribution unit is no longer a single author, but the group - institutions or country. Cronin pointed out that it is very difficult to assess the contribution of a person to an
article under this phenomenon, as we cannot figure out which author is dominant. Now for "hyperauthorship" article, credit is generally based on the average allocation, such as SCOAP3’s method (Krause J & Lindqvist C M & Mele S, 2007). But this distribution is also biased. Birnholtz J P (2006) stated that there are many authors in the "hyperauthorship" articles but are not involved in specific work in the researchers' interview, in this case, give the same credit to all people is obviously unfair. So it needs to explore a new way to measure the contribution in the case of "hyperauthorship".

This issue keeps undetermined currently. Some studies heuristically mapped the contribution of authors to their signature order. Yet there is no agreement on the signature rule in different disciplines or even in one discipline. Some used the citing relation to evaluate the author. But not all the citation is valid for the discussed article. Some highlighted the credit of the authors of an article who had publications on similar subjects. Yet these publications were usually located by the citing articles of the discussed article. If the citing articles may not thoroughly cover the similar publications, or the discussed article was newly published, the method would not work well. So how to find a more reasonable and universal method to allocate the authors’ credit is discussed in the following.

Related work

In the study of credit allocation, two main factors are considered and discussed.

The signature order of author

(1)Some scholars have pointed out that the authors' signatures are now usually arranged in terms of the contribution to the research. This view has been proposed and elaborated by Hodge and Greenberg (1981). Van Praag (2007) elaborated on the benefits of the author who is the first in signature, so people are more willing to put the author of the greatest contribution to the first one. Ludo Waltman (2012) has proved that the majority of the authors are willing to decide the signature order depending on the contribution of authors and authors’ number with such rule presenting an increasing trend, combined with the practical experience and the actual statistics. In addition, there are many algorithms which are based on this rule: Proportional Counting, which was proposed by Hooydonk (1997), in which the credit of each author decline in equal difference by the order; later, Abbas (2011) proposed another method, and this algorithm is more flexible than the Proportional Counting because it adds a mediation parameter that can adjust the decreasing disparity. Such linearly algorithms have its inevitable flaw, for the same order r of the author, the credit will increase first and then decrease with the number of authors increase, which is clearly not consistent with the actual experience. In addition, there are some nonlinear algorithm: Eghe (2000) put forward the Geometric Counting, he believed that this model is more in line with the actual situation than the Arithmetic method; Hagen (2010) proposed Harmonic Counting, which was verified to be the most practical algorithm (without adjustable parameters).

(2)In addition to the rule that signature is ordered by contribution, some scholars (Hodge S E, 1996; Huang M H, 2011; Tove Faber Frandsen & Jeppe Nicolaisen, 2010) mentioned that signature with alphabetical order is also common. This rule is more prominent in certain fields, such as the economy (including finance), high-energy physics and other fields (Ludo Waltman, 2012). Based on this point of view, some of the classical methods are Total Counting and Fractional Counting (Lindsey D, 1980), the former method will cause a significant expansion problem.

(3) In addition, there are many methods which don’t adopt the above two theoretical assumptions. For example, First-author counting, which takes into account only the contribution of the first author and allocates all the credit to the first author, which is clearly unfair to other authors. In CCS (Lukovits I& Vinkler P, 1995), Trueba & Guerrero (2004), and Zhang (2009), the first author and the corresponding author are calculated separately and considered to be same important.
This article has summarized several algorithms based on signature orders, which are often discussed at present, see Table 1:

**Table 1. Algorithms for credit allocation based on signature orders.**

<table>
<thead>
<tr>
<th>Algorithm’s name</th>
<th>Mathematical formula</th>
<th>Algorithm’s Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Counting</td>
<td>( C_r = 1, 1 \leq r \leq N )</td>
<td>Not consider the number of authors. Credit is expansion.</td>
</tr>
<tr>
<td>Fractional Counting</td>
<td>( C_r = \frac{1}{N}, 1 \leq r \leq N )</td>
<td>When N is large, the contribution of all authors is close to zero.</td>
</tr>
<tr>
<td>Proportional Counting</td>
<td>( C_r = 2 \times \left( 1 - \frac{r}{N+1} \right), 1 \leq r \leq N )</td>
<td>Credit increases linearly with the order of authors’ signature. The credit of the same order is non-monotonous as the number of co-authors increases.</td>
</tr>
<tr>
<td>Arithmetic Counting</td>
<td>( C_r = \frac{N-r-2}{N} \times \alpha, 1 \leq r \leq N, 0 \leq \alpha \leq 1 )</td>
<td>Credit increases linearly with the order of authors’ signature. The credit change as a curve line with the number of co-authors increases. Parameters are adjustable, used to control the difference between the credits of the author. The first author’s credit with the increase of the number of participants decreases very rapid.</td>
</tr>
<tr>
<td>Geometric Counting</td>
<td>( C_r = \frac{2^{N-r} - 1}{2^N - 1}, 1 \leq r \leq N )</td>
<td>The credit change as a curve line with the number of co-authors increases. A few authors get the most contribution.</td>
</tr>
<tr>
<td>Sekercioğlu Schema</td>
<td>( C_r = \frac{1}{r}, 1 \leq r \leq N )</td>
<td>The credit change as a curve line with the number of co-authors increases. The overall credit is high. The credit between the authors is quite different, especially in the former author.</td>
</tr>
<tr>
<td>Harmonic Counting</td>
<td>( C_r = \frac{1}{\sum_{i=1}^{N} r^{-i}}, 1 \leq r \leq N )</td>
<td>The credit change as a curve line with the number of co-authors increases. The degree of matching is high in the practical assessment.</td>
</tr>
<tr>
<td>First-author Counting</td>
<td>( C_r = \begin{cases} 1 \text{,} &amp; r = 1 \ 0 \text{,} &amp; 1 &lt; r \leq N \end{cases} )</td>
<td>Only consider the first author.</td>
</tr>
<tr>
<td>SDC Schema</td>
<td>( C_r = \begin{cases} 1 \text{,} &amp; r \leq 10 \ 0.95 \text{,} &amp; r &gt; 10 \end{cases} )</td>
<td>Be suitable to the article with a large number of authors. Focus on the top 10 authors.</td>
</tr>
</tbody>
</table>

**The authority of authors**

The authority of an author in one research field can determine whether he or she has the capability to contribute to the target research. The indicators include the author's output in this field and the citations the target article gained:

1. The output of the author is a very straightforward measurement. Kosmulski M (2012) believe that the higher the output is, the more qualified the author can be. But the rule is unfair to authors who were not productive yet had high quality papers.

2. Compared with the simplicity of the output metrics, citation reflects the impact of an article. Garfield (1972) initially counted the citation for the evaluation of articles or researchers. Later, h-index (Hirsch, 2005) was put forth, which is now quoted in some literature database, such as Web of Science, Scopus, etc. The variants of h-index includes g-index (Egghe L, 2006), c-index (Bras-Amorós M, 2011), s-index (Silagadze Z K, 2009), etc. Besides, there is an algorithm (Pradhan D, 2016) similar to PageRank, the citation by other authors in this field can indirectly reflect author's authority. However, the above method is
more suitable for the evaluation of the authors’ scientific achievements in the whole domain, not in a single article. Wang H, Shen HW, Cheng XQ et al. (2016) pointed out that the authors' credit allocation of an article could be considered in two aspects: the author network and the article network. The difference of them is that an article will be given corresponding gross value according to citation by article network, and equal distribution by author network. They argued that using author network may lead to misallocation of contributions, especially for the authors whose articles are high quality but small quantity.

(3) The co-citation is the extension of the unidirectional citation relation. The utility of co-citation has been mentioned early (Garfield E, 1972; White H D & Mccain K W, 1998), which could be the tools of journal evaluation, or find the experienced author in some domains, and so on. Author co-citation analysis can be used in lots of research, author assessment is one of that, Ding Y & Yan E. (2009) used the co-citation network to rank the authors. Hua-Wei Shen et al. (2014) identified the highest credit author using the co-citation of Nobel candidate articles, the special part of which is the assessment in a single article, and the combination with the methods based on the signature order. For convenience, we call it co-cited model.

**Model based on co-cited articles**

**Model description**

The co-cited model was proposed by Hua-Wei S et al. (2014). In this model, one author will be given a basic value basing on Fractional Counting in the discussed article, then the final value will be the sum of the credits in the co-cited articles of this author. Finally allocate the credit depending on the proportion of each author’s final value. They validate the method by identifying the authors of Nobel-winning articles that are credited for the discovery. The method’s accuracy achieved about 75%.

**Model analysis of advantages and disadvantages**

(1)The advantage of using co-cited articles: firstly, co-citation can aggregate the literature with similar content; secondly, it can expand the time of relevant document, reference or citation can only appear on the unilateral side of the target document (before the target document or after), and the co-citations can cover the entire time axis.

(2)The co-cited model adopted Fractional Counting as the basic credit calculation. Although this algorithm can reduce the burden of calculation to a certain extent and avoid the extremely uneven distribution when it cannot judge the ranking intention. So how to adopt fit algorithm needs further consideration.

(3)The co-cited articles play a role of the aggregation of similar content literature. But the model only uses its co-citation "Frequency", apart from that, is there any other properties can be used to measure the importance?

(4)The primary condition of co-cited model is that there are a certain amount of citations. When the number of citations of the target literature is insufficient, the model cannot work well. So for the literature with less co-citations, what methods can make up for this defect?

**The improvement to co-cited model**

Based on the articles with different range of authors' number in the field of high energy physics, this research developed the co-cited model, and compare with the original co-cited model and several algorithms based on signature order, for convenience, algorithms solely based on signature order can be called as traditional methods.

The main research questions are as follows:

**Question-1** How should the order of authors' signatures be applied to the allocation of contributions, where the articles have different number of co-authors?

**Question-2** How to measure the author's authority over the content? Is there any other ways more reasonable than the original citation model?
Exploration of authors’ signature rules

In order to explore the relationship of authors' signature and the number of authors, this article analyzed 10046 papers from the SCOAP3 open access platform. The data involved a total of 61513 authors, and the largest number of co-authors is 2951 in a single article.

This article compares the real signature order with the simulated order generated by other permutations, like author's alphabetical order, country's alphabetical order, etc. If it is consistent with the real order, it is more likely to be arranged in accordance with the rule. But, the more the number of co-authors, the higher the likelihood of anastomosis:

1. The number of co-authors less than 5: for these articles, most authors’ institutions and countries are the same, and the articles in accordance with the author's alphabetical order take up 72.91% (except for only one author's article). But it may be accidental due to the small number of authors. In this case, the signature law is more difficult to be reflected from the data. It is relatively clear in task allocation, and the gap between the authors should not be so obvious.

2. The number of co-authors is between 5 and 10: 13.21% of the articles that meet the national alphabetical order; 54.19% meet the alphabetical order of the name. In this case, the rule is more pronounced, and the articles by alphabetical order and none is close in number, so we must judge the order rule first and then decide the distribution method. For such literature, leading contributors are relatively significant but not too much, and the gap between non-leading contributors should not be too large.

3. The number of co-authors greater than 10: these articles almost follow the order of a certain type of alphabet, for example, the national order, the name order, the organization order and so on. In this case, the law is clear and single, basically alphabetical order. Because the number of people is too high, this law can be explained. The uniform distribution is a more reasonable algorithm for them in traditional algorithms.

According to the signature rule, the three types of literature are suitable for different traditional algorithms. The rationality is weak if solely depends on signature rules, so it should be supplemented with content correlation.

Principle of model improvement

This article have analyzed the merits and demerits of the original co-cited model. Therefore, our study maintains the advantages of co-cited model and improves its shortcomings. There are still some problems in current methods, the authors proposed some solutions to the following phenomena:

**Phenomena-1** “Hyperauthorship”.

Problem The contribution can’t be allocated to individuals.

Solution Increase the weight of the target author according to co-cited articles; compare the content of the article with the author's research experience.

Basis The phenomenon changes the research mode of cooperation from individuals-individuals to groups-groups; the traditional methods are based on empirical rules, which is not suitable for all situations.

**Phenomena-2** The diversity of signature rules in the articles with medium number of co-authors.

Problem This kind of articles cannot be treated with one method.

Solution In the basic contribution calculation, to judge whether the order of the signature follows the author's name, if it is, adopt the Fractional Counting, otherwise adopt the Harmonic algorithm.

Basis Features of such articles are obvious, easier to judge. (Seen in “Exploration of authors' signature rules”); Harmonic Counting has been proved to be the most realistic in the traditional algorithms without adjustable parameters.

**Phenomena-3** Uncertainty in the signature rules of minor authors' articles.
**Problem** It is difficult to determine the actual rule of signature by the order of the author.

**Solution** In the basic contribution calculation, adopt the Proportional Counting, a typical linearly decreasing algorithm.

**Basis** The contribution and task distribution is rather clear in these articles; most scholars agree that the signature is sorted by the contribution from high to low; compared with the linearly decreasing algorithm, the gap between the authors in the nonlinear algorithm is relatively large, which is not suitable for too few and too many authors.

**Phenomena-4** Not all the citation relationship is valid for the target article.

**Problem** Some citation which has nothing to do with the target content will be included, which led to the allocated contribution in the current literature is not fair and reasonable.

**Solution** After choosing the related articles by co-citation firstly, compare the target content and the authors’ research experience.

**Basis** The citation can aggregate articles roughly, and the content match will make a more detailed distinction between the articles.

**Phenomena-5** Most of the articles have less citations.

**Problem** The total amount being cited is less, or even zero, making the co-cited model’s effectiveness cannot be achieved.

**Solution** On the basis of the co-citation, consider the author’s self-citing relationship.

**Basis** If the other study of the target author can provide reference for the discussed study, the author should be more likely to make a greater contribution to the target article; the number of citations is not influenced by the publication time of the article or the quality of the article, so the source of the reference is more stable for all documents.

**Data collection and description**

The data used in this study was from the Web of Science database, the authors first selected two journals in HEP field, whose article number is large, from the journals obtained by SCOAP3(https://scoap3.org). They are Journal of High Energy Physics and Physics Letters B.

Considering the use of citation data and time lag of citing, so the authors selected the papers with citation times more than 150, published during 2000-2010. The sample capacity of 528, then according to the number of co-authors stratified sampled the seed literature, among which there are 20 for 1 to 4 co-authors, 14 for 5 to 10, 11 for 10 +.

This study involved four sets of data: the seed literature, the citation of the seed literature, the co-citation of the seed literature, and the reference literature of the seed literature. There are 1367 authors in seed sets, and the total citation of seed literature is 14080 times; 973,185 times co-citations, in which there are 105,880 articles.

The downloaded data was involved in the nine fields and explained in Table 2:

**Table 2. The field identifier of WOS data.**

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Description</th>
<th>Identifier</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF</td>
<td>Full name of the author (by signature)</td>
<td>C1</td>
<td>Address of authors</td>
</tr>
<tr>
<td>TI</td>
<td>Title</td>
<td>RP</td>
<td>Address of corresponding author</td>
</tr>
<tr>
<td>DE</td>
<td>Keywords</td>
<td>CR</td>
<td>Reference (with DOI number)</td>
</tr>
<tr>
<td>ID</td>
<td>Additional keywords</td>
<td>DI</td>
<td>Digital object identifier (DOI)</td>
</tr>
<tr>
<td>AB</td>
<td>Abstract</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DI is the unique identifier of an article, and CR is used to obtain the citation literature of the seed literature. AF and C1 can show the order of the authors’ signature, and the attribution
of organization and nationality(state), and the RP points out the corresponding author, besides, the content of the article can be explained by TI, DE, ID, AB.

This study has used RegexBuddy 4, java programming to preprocess the data. RegexBuddy is widely used for regex expression editing, which is combined with java for the extraction of the required information. Besides, if ID and DE are both missing, from the TI field and AB field to obtain the representative words using TextRank.

The implementation steps of the improved model

(1)First part: The basic contribution’s superposition basing on the co-citation and self-citation.

This part is an improvement to the basic credit of the original co-cited model, and take the self-cited relationship into consideration.

The improved model involves four types of literature sets: T-the seed articles; D-the articles citing T; G-the articles co-citing T; F-the articles cited by T. And the basic term symbolic keep consistent with original co-cited model.

∀t ∈ T:

\[ M_t \] is the set of all authors in t, |\( M_t \)| = \( p \), for \( p_0 \), \( M_{p_0} = \{a_1, a_2, a_3, a_4\} \), |\( M_{p_0} \)| = 4; \( D_t \) is the set of articles citing t; \( G_t \), the terms of which have the similar content with s; \( P_t = D_t \cup G_t \cup F_t \), |\( P_t \)| = q, all literature of \( P_t \) determine the weight of authors in \( M_t \).

Step1 Count citation times of \( P_t \) by \( G_t \), then get s.

Step2 Calculate the basis credit in a single article. The following logical judgments are made for the literature in the \( P_t \), for \( p_0 \):

IF \( \forall p_i \in P_t \) , the author of \( p_i \) is one of the author in \( M_{p_0} \), \( a_k \in M_{p_0} \), r is the order, N is the number of authors in \( p_i \).

IF \( 1 \leq N \leq 4 \)

\[ A_{a_kpi} = C_r = \frac{2 \times (1 - \frac{r}{N+1})}{N}, 1 \leq r \leq N \]

ELSE IF \( 5 \leq N \leq 10 \)

IF signature is ordered by alphabetic

\[ A_{a_kpi} = C_r = \frac{1}{N^r}, 1 \leq r \leq N \]

ELSE

\[ A_{a_kpi} = C_r = \frac{1}{\sum_{i=1}^{N} r}, 1 \leq r \leq N \]

ELSE

\[ A_{a_kpi} = C_r = \frac{1}{N}, 1 \leq r \leq N \]

ELSE

\[ A_{a_kpi} = 0 \]

As for corresponding author, the credit of which should not affect others’ credit, if the basic algorithm is average allocation, the credit of corresponding is double of original, else the credit directly equivalent to the one ranked second (if the first author is corresponding, the credit maintains).

Step3 \( c = A \cdot s \), c is the proportion of credit allocated to every author in \( p_0 \).

(2)Second part: The content matching based on co-citation and self-citation.

This part is to compare the similarity of author’ research experience and the content of target literature. Then according to the similarity allocate the credit:
Step 4 For \( p_{0} \), the set of keywords is \( d_{p_{0}} \), \( a_{k} \in M_{p_{0}} \), the keywords set of \( a_{k} \) is \( b_{a_{k}} \), the keywords of which can be got from \( P_{p_{0}} \);

Step 5 To compress \( b_{a_{k}} \). In order to reduce the cost of the program processing, compress should be carried out before the similarity calculation. The purpose is to remove the similar words in the set, so that the words in the collection cannot replace each other. If \( |b_{a_{k}}| > 50 \), do the compress, \( MIN_{sim} \) was set to be 0.6 at first, the first term of \( b_{a_{k}} \) was added to set \( f_{a_{k}} \), then compare the next term with all the terms of \( f_{a_{k}} \), if one of the similarity of \( b_{a_{k}}[i] > MIN_{sim} \), then add sim to the weight of \( b_{a_{k}}[i] \). If the \( |f_{a_{k}}| > 50 \) after a round, then set \( MIN_{sim} = MIN_{sim} - 0.1 \) and do compress again.

Step 6 To compare the terms of \( f_{a_{k}} \) and \( d_{p_{0}} \), then sum the similarity \( SIM_{a_{k}} \), for \( p_{0} \) \( g_{p_{0}} = \langle SIM_{a_{1}}, SIM_{a_{2}}, SIM_{a_{3}}, SIM_{a_{4}} \rangle \), that is the proportion of allocation.

In this article, the method used in the comparison is based on the semantic similarity of WordNet2.1 dictionary.

(3) Third part: Get the final contribution credit.

This part is to sum the score of former two parts, and get the final credit.

Step 7 for \( p_{0} \), the final credit proportion of \( a_{1}: a_{2}: a_{3}: a_{4} = (c_{p_{0}}[1] + g_{p_{0}}[1]): (c_{p_{0}}[2] + g_{p_{0}}[2]): (c_{p_{0}}[3] + g_{p_{0}}[3]): (c_{p_{0}}[4] + g_{p_{0}}[4]) \).

Result of state contribution based on the improved model

Summing up the authors’ contributions by the state attribution of each author. In order to compare these methods fairly later, the authors standardized the original state contribution value: \( s_{a_{k}} = \frac{MIN_{a_{k}}}{MAX_{a_{k}} - MIN_{a_{k}}} \). See Table 3 for details:

<table>
<thead>
<tr>
<th>Rank</th>
<th>State</th>
<th>Contribution</th>
<th>Standardized contribution</th>
<th>Rank</th>
<th>State</th>
<th>Contribution</th>
<th>Standardized contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>USA</td>
<td>8.215825782</td>
<td>1</td>
<td>20</td>
<td>Hong Kong</td>
<td>0.190191757</td>
<td>0.022754037</td>
</tr>
<tr>
<td>2</td>
<td>Italy</td>
<td>4.968350191</td>
<td>0.604569259</td>
<td>21</td>
<td>Russia</td>
<td>0.187835065</td>
<td>0.022467073</td>
</tr>
<tr>
<td>3</td>
<td>Germany</td>
<td>3.92423109</td>
<td>0.477431493</td>
<td>22</td>
<td>Poland</td>
<td>0.183189726</td>
<td>0.021901431</td>
</tr>
<tr>
<td>4</td>
<td>China</td>
<td>1.852965957</td>
<td>0.225226968</td>
<td>23</td>
<td>Portugal</td>
<td>0.069171932</td>
<td>0.008017988</td>
</tr>
<tr>
<td>5</td>
<td>Japan</td>
<td>1.644148636</td>
<td>0.199795936</td>
<td>24</td>
<td>Finland</td>
<td>0.052644282</td>
<td>0.006005489</td>
</tr>
<tr>
<td>6</td>
<td>Taiwan</td>
<td>1.620630503</td>
<td>0.196932237</td>
<td>25</td>
<td>Armenia</td>
<td>0.043114895</td>
<td>0.004845138</td>
</tr>
<tr>
<td>7</td>
<td>France</td>
<td>1.15546309</td>
<td>0.140290858</td>
<td>26</td>
<td>Netherlands</td>
<td>0.039974335</td>
<td>0.004462726</td>
</tr>
<tr>
<td>8</td>
<td>CERN</td>
<td>1.092924327</td>
<td>0.13267579</td>
<td>27</td>
<td>Czech Republic</td>
<td>0.038393317</td>
<td>0.004270212</td>
</tr>
<tr>
<td>9</td>
<td>Hungary</td>
<td>0.980542325</td>
<td>0.11899153</td>
<td>28</td>
<td>Mexico</td>
<td>0.035651248</td>
<td>0.003936323</td>
</tr>
<tr>
<td>10</td>
<td>England</td>
<td>0.86246874</td>
<td>0.104614232</td>
<td>29</td>
<td>Romania</td>
<td>0.029429082</td>
<td>0.003178677</td>
</tr>
<tr>
<td>11</td>
<td>Spain</td>
<td>0.759622003</td>
<td>0.09209104</td>
<td>30</td>
<td>Australia</td>
<td>0.021055771</td>
<td>0.002159096</td>
</tr>
<tr>
<td>12</td>
<td>UK</td>
<td>0.733104164</td>
<td>0.08886208</td>
<td>31</td>
<td>Belgium</td>
<td>0.015791881</td>
<td>0.001518135</td>
</tr>
<tr>
<td>13</td>
<td>India</td>
<td>0.650925944</td>
<td>0.078855601</td>
<td>32</td>
<td>Slovenia</td>
<td>0.013474423</td>
<td>0.001235949</td>
</tr>
<tr>
<td>14</td>
<td>Greece</td>
<td>0.543245168</td>
<td>0.065743789</td>
<td>33</td>
<td>Vietnam</td>
<td>0.006380874</td>
<td>0.000372199</td>
</tr>
<tr>
<td>15</td>
<td>Brazil</td>
<td>0.386390239</td>
<td>0.046644258</td>
<td>34</td>
<td>South Korea</td>
<td>0.004772179</td>
<td>0.000176315</td>
</tr>
<tr>
<td>16</td>
<td>Thailand</td>
<td>0.358500142</td>
<td>0.043248205</td>
<td>35</td>
<td>Croatia</td>
<td>0.004726053</td>
<td>0.000176968</td>
</tr>
<tr>
<td>17</td>
<td>Canada</td>
<td>0.356046659</td>
<td>0.042949455</td>
<td>36</td>
<td>South Africa</td>
<td>0.004654798</td>
<td>0.000162022</td>
</tr>
</tbody>
</table>
Argentina 0.345975816 0.041723173 37 Peru 0.003324192 0

*Notes: The authors treated CERN as a state (Krause J & Lindqvist C M & Mele S, 2007); for an author with multiple state attribution, all the attribution were given the same contribution value.

Table 4. The comparison of several methods correlation (spearman correlation).

<table>
<thead>
<tr>
<th></th>
<th>Proportional</th>
<th>Harmonic</th>
<th>Fractional</th>
<th>Segmented algorithm</th>
<th>Co-cited model</th>
<th>Improved model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportional</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harmonic</td>
<td>0.987</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fractional</td>
<td>0.980</td>
<td>0.965</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Segmented</td>
<td>0.985</td>
<td>0.971</td>
<td>0.993</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>algorithm</td>
<td>Co-cited</td>
<td>0.960</td>
<td>0.948</td>
<td>0.973</td>
<td>0.968</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improved</td>
<td>0.872</td>
<td>0.867</td>
<td>0.894</td>
<td>0.890</td>
<td>0.921</td>
<td>1</td>
</tr>
<tr>
<td>model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The original co-cited model takes the Fractional Counting and co-cited articles into consideration; the improved model considers the self-citation and co-citation relations in the calculation of the basic contribution, and segmented traditional algorithm, besides, it introduces the thought of content match. Fractional Counting, Harmonic Counting, and Proportional Counting are three typical pure author signature algorithms. The segmented traditional algorithm is a combination of the fractional counting, harmonic Counting,
proportional Counting, which classifies the basic contribution calculation under different situations.

In order to compare the similarity of the results of these algorithms, the authors calculated the national contribution based on different algorithms, and uses the spearman correlation coefficient to measure (using R language).

According to Table 4, the correlation between the new model and several other methods is relatively low, of which thought has a relatively large impact on the results of the national contribution calculation, especially relative to the four traditional algorithms; The results of the Fractional Counting present a line parallel to the x-axis, the Proportional Counting is linearly decreasing line, and the Harmonic Counting is a curve line. The results of the authors’ contributions are different in a single article. However, the correlation between the three algorithms is higher than 96%, which show that although the choice of traditional algorithms is different, the impact to the contribution of states is relatively small, and the impact of the segmentation of traditional algorithm is small too; In general, the difference of the algorithms based on the signature order would be diluted in the process of states’ contribution calculation, but the inclusion of citation and content match can have a greater impact.

(2)Single factor analysis of the contribution to individual authors

1)Self - citation

According to statistics, the target literature involved a total of 1367 authors, the number of self-cited is 2709 times, and an average of 1.98 per author, that is, self-cited relationship can bring the superposition of two articles’ contribution to each author for average, which takes an increase in the original contribution of 36.71 (not a standardized contribution to each document), compare to the pure co-cited model, it brings a total of 5% contribution credit. In conclusion, just refer to the number and additional contribution of self-citation, the self-citation can make up for the lack of co-citation, but in our experiment, because the co-citation of sample is too many, so the effect of self-citation is not obvious.

2)Co-citation

In order to verify the influence of this factor on the allocation of individual authors, the t test (using SPSS) of pair of samples was used to calculate the contribution of 500 random authors based on the results of fractional counting and original co-cited model. The results of the test showed that p = 0.053 <0.1, which shows that co-citation factors have a significant effect on the authors’ distribution of individual documents at a confidence level of 90%.

3) Classification of Three Classification Basic Algorithms

The basic contribution algorithm in the new model is based on the different algorithms of the sub-case. We also select 500 random authors as samples, and set two control variables in the test group (using SPSS): (A) Fractional + Proportional + Harmonic, (B) Fractional + Proportional and (C) Fractional + Harmonic. The test results show that the p value of (A) vs (B) is 0.687, and the p value of (A) vs (C) is 0.998. Therefore, the participation of Proportional and Harmonic algorithms has little effect on the contribution credit of single literature.

4) Research content matching

500 random authors are selected. These two parts contribute to the distribution of the results of the t-test of the paired samples (using SPSS):
Table 5. The t-test of the paired sample in the improved model.

<table>
<thead>
<tr>
<th>Experimental group and control group</th>
<th>Controlled variable</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>The improved model(including two parts) VS The improved model(basic contribution superimposed)</td>
<td>The content matching</td>
<td>0.017</td>
</tr>
<tr>
<td>The improved model(including two parts) VS The improved model(content matching)</td>
<td>The basic contribution superimposed</td>
<td>1.000</td>
</tr>
</tbody>
</table>

It can be seen from Table 5 that for the two parts of the improved model, p-value of the content matching factor is less than 0.03, that is, at a confidence level of 97%, content matching can have a very significant effect on the final results of the model. From the comparison of the p-values of the two parts, it can be seen that in the improved model, the content matching factors play a more important role than the basic contribution calculation.

### Conclusion and future work

In this article, 45 articles of high-energy physics field were selected by stratified sampling as the seed set, and the self-cited, citing and co-cited articles were obtained for the seed set. The authors have improved original co-cited model from the following two aspects: first, considering the self-citation and co-citation relations in the calculation of the basic contribution, and classifying the basic contribution calculation under different situations; the second is comparing the research direction of the authors and the subject content of the target document. Finally comparing the similarity and effect of methods.

From the comparison of the methods, each pair of contrast is more than 0.8, we can see that the rationality of these algorithms can be mutually confirmed. The spearman values between the improved model and other methods are the smallest, the impact of the improvement is relatively large; different types of traditional algorithms perform different in a single document, but in the state contribution the similarity of the results is more than 0.96, whose difference is small. Therefore, when the algorithm of national contribution needs to be improved, the effect of traditional algorithm change is not obvious, so it should be more effective in exploring the citation relationship and research content.

As for the single factor analysis: (1) self-citation can bring 1.98 articles to strengthen the author weight in the target article for average. If the article has many co-cited articles, the effect of self-citation would be diluted. But it can make up for the lack of co-citation. (2) The introduction. The matching sample t test shows that the co-citation will have a significant impact on the allocation of authors' contribution at a confidence level of 90%. (3) All traditional algorithms have no obvious effect. In general, the improvement of the algorithm depending the rules of authors' signature is not good, whether for the contribution of state or the single author. (4) Research content matching has a very significant effect on the distribution of individual authors' contributions in the new model, at a confidence level of 97%. The effect of this factor is greater than that the basic contribution calculation.

In conclusion, the experimental results show that the influence of the traditional algorithm on the final effect is not obvious, and the introduction of the common reference is a significant improvement to the traditional algorithm. The self-citation relationship can make up for some literatures. The matching of the research contents has a very significant effect on the author's contribution allocation algorithm. The improved model is able to reflect its own characteristics under the premise of ensuring the rationality of the general trend.

There are some shortcomings in this study hope that future research can be improved: first, to expand the amount of sample data, which can more effectively illustrate the authenticity of the results; second, the universal test in other fields, not limited to the field of high-energy physics; Thirdly, in the research content matching, word vector can be combined to calculate the similarity of the text in the field, so that the matching is more accurate and
professional; Lastly, in this article we did not compare the reality, that can be the assessment of professors in HEP, or the origin authors’ self-evaluation. It’s really a pity for the lack of that, hope the future work will rich the study to validate the model’s effectiveness more comprehensively.

Acknowledgments

Thanks Institute of Scientific and Technical Information of China (ISTIC) for providing SCOAP3 data, which was used in the exploration of authors’ signature rules. And the authors are sincerely grateful for the anonymous reviewers’ advices during the process of paper writing.

References


Garfield, E. (1972). Citation analysis as a tool in journal evaluation. Science, 178(4060), 471.


Hu, X. (2009). Loads of special authorship functions: linear growth in the percentage of "equal first authors" and corresponding authors. Journal of the Association for Information Science and Technology, 60(11), 2378-2381.

Hu, X., Rousseau, R., & Chen, J. (2010). In those fields where multiple authorship is the rule, the h-index should be supplemented by role-based h-indices. Journal of Information Science, 36(1), 73-85.


Li Z, Sun YM, Wu FX, Yang LQ, Lu ZJ, & Yu WF. (2013). Equal contributions and credit: an emerging trend in the characterization of authorship in major anaesthesia journals during a 10-yr period. Plos One,8(8), e71430.


Do Patent Citations to Conference Papers Differ from Journal Articles? Evidence of Technological Impact through Google Patents

Ashraf Maleki

Malekiashraf@ut.ac.ir
Master Graduate at Scientometrics, University of Tehran, Iran

Abstract
The research presented in scientific meetings are a major way to maintain the connection between science and technology in most engineering-related fields. However, although conference papers constitute predominant proportion of publications in some disciplines, the extent to which they are able to be cited in patents compared to journal articles is less known. In order to examine this, bibliographic information of a sample of 111,108 Scopus indexed papers including journal articles (53%) and conference papers (47%) across 30 disciplines in two years (1996 and 2006) was searched in Google Patents (GP) using Bing search engine. An ordinary least squares regression model demonstrated that conference papers and journal articles were almost equally predictive of patent uptakes in most fields, suggesting that proceeding can be as important as journal articles if technological value is meant to be assessed. Although patent uptake of both journal articles (2.9%) and conference papers (1.9%) was relatively low, disciplinary differences should also be considered since journal articles are acting better in fields such as biomedical engineering and bioengineering, but conference papers are significantly better in geotechnical engineering, automotive engineering and computer science applications. Furthermore, it takes slightly more years for conference papers than journal articles to be cited in patents.

Conference Topic
Patent Citation Analysis, Altmetrics, conference papers, technological impact

Introduction
Conference proceedings tend to be premier venue when publishing innovative knowledge in line with trendy topics (Chen, Song, & Zhu, 2007) or printing theory-changing breakthroughs (Winnink, Tijssen, & van Raan, 2015). However, due to inconsistent acceptance rate and precision attributed to peer review processes in some scientific meetings (Cabanac & Preuss, 2013), they have been relatively undermined in research impact assessment practice compared to formal journal papers. Both Web of Science and Scopus have improved their conference papers’ citation indexes (http://wokinfo.com/products_tools/multidisciplinary/webofscience; Scopus Content Coverage Guide, 2016), although their coverage might be inconsistent across fields (Bar-Ilan, 2010; Meho & Yang, 2007). However, scholarly citations to conference papers are not prevalent (Michels & Fu, 2014), partly due to authors tendency to cite equivalent articles in periodicals (Lisée, Larivière, & Archambault, 2008). On the other hand, it is also argued that conventional citations might fail to demonstrate actual value associated with proceedings because conference paper are sometimes intended to trigger formal publications (Drott, 1995) and aid scholarly communications unlike the way journal articles do. Some studies presented at scientific meetings maintain the connection between knowledge and application for example in biotechnology (Martens, 1993), or the interaction between industrial and academic researchers such as in microelectronics and software engineering (Meyer-Krahmer & Schmoch, 1998); they are also a significant output to measure productivity of science-based technical projects (Breschi & Malerba, 2011; Garg, Gupta, Jamal, Roy, & Kumar, 2005). In consequence, conference papers are sometimes recognised with more ability to measure technical innovation and novel ideas rather than science production (Montesi & Owen, 2008). However, despite the proposed bibliometric indicators for assessing quality of scientific meetings (Martins,
Gonçalves, Laender, & Ziviani, 2009), a measure to recognise likely technological impact of conference papers is less investigated.

Some bibliometric investigations into proceeding publications tried to identify research topics (Hofer, Smejkal, Bilgin, & Wuehrer, 2010) and potential collaborations in science (Boyack, 2009). Online indicators (altmetrics) was also analysed to assess conference papers’ Mendeley readerships compared with journal articles in engineering-related fields (Aduku, Thelwall, & Kousha, 2017). Furthermore, citations in online conference presentation files was used as evidence of research impact (Thelwall & Kousha, 2008) and tweets posted during conference events were also used to assess scholarly communications in scientific meetings (Weller, Dröge, & Puschmann, 2011). However, few have analysed patent uptake of conference papers as evidence of their technological impact and there has been no multidisciplinary evaluation compared with journal articles. This article partly fills the gap by speculating patent citations to both conference papers and journal articles in 30 scientific disciplines mainly constituting from engineering-related disciplines and some other fields with considerable proportion of literature emerged in conference proceedings. The analysis is conducted using webometric method and through Google Patents searches.

Scholarly Impact of Conference Papers

Given the importance of industry and technological connections in conferences, assessing technological impact of conference papers seems to be important, because in some engineering-related disciplines scientific meetings are leading form of scholarly communications (Goodrum, McCain, Lawrence, & Giles, 2001). A study of ISI proceedings in 1997 through 2005 showed that conference papers constituted up to around 55% of publications in engineering (Glänzel, Schlemmer, Schubert, & Thijs, 2006) and over 70% of 2009 publications in electrical engineering and computers (Michels & Fu, 2014). Some countries such as China had significant growth in terms of number of conference papers in computer sciences, though the growth in citation rate tended to be minimal (He & Guan, 2008). As various studies have pointed out, significant proportion of publications in computer sciences are conference papers (Bar-Ilan, 2010; Vrettas & Sanderson, 2015), and assessing their technological impact perhaps can make sense because some conference papers recognized with leading discoveries in computer science were witnessed without formal citations available (Winnink et al., 2015). This poor reflection of impact might be due to partial coverage in conference citation indexes (Meho & Yang, 2007) as well as different nature of their impact (Drott, 1995; Lisée et al., 2008). However, the extent to which conference papers tend to demonstrate technological value compared to formal journal articles is less known. To address this issue, current research aimed at examining technological value associated with conference papers through patent citations as an alternative indicator of research assessment (Wouters et al., 2015).

Patent impact of Academic Research

Numerous studies have discussed commercial value and technological utility associated with public science through evaluating patent citations to academic research (Meyer, 2000; Narin & Olivastro, 1998; Oppenheim, 2000; Schmoch, 1993), as journal and conference references are found as major constituents of non-patent references. One study on two samples of references in granted patents from 1991 to 2001 estimated proportions of citations to conferences and journals as much as 17% and 55%, respectively, in USPTO, and 34% and 64% in EPO (Callaert, Looy, Verbeek, Debackere, & Thijs, 2006). There is also a lower estimation of conference paper references (5.9% vs. 77.7%) for USPTO Chinese patents in nanotechnology between 1991 and
2008 (Wang & Guan, 2011). Another research did a large-scale estimation of patent uptake for Scopus indexed journal articles between 1996 and 2012 using Bing’s Google Patent search results, finding patent uptake of articles not prevalent, ranging between a mere 3.2% in mechanical engineering and approximately 10% in biomedical engineering (Kousha & Thelwall, 2017). A webometric study recognised that extent of Google Patent (GP) indexed URL references outlinking to academia was able to demonstrate technological influence of US universities (Orduna-Malea, Thelwall, & Kousha, in press). Furthermore, journal articles citations in patents seem to be able to recognise a distinct aspect of research impact relative to conventional citations (Sud & Thelwall, 2014) because they are found in very weak correlations with each other (Kousha & Thelwall, 2017; Tijssen, Buter, & van Leeuwen, 2000).

Research Question

In order to be able to identify whether a particular document type advantages more patent citation both conference papers and journal articles are assessed and compared in relation to Scopus citations and over time. Therefore, current research is driven by below questions:

1. To what extent do conference papers have patent impact in relation to journal articles?
2. In which fields are there disproportionate patent citations to conference papers or journal articles?
3. To what extent patent citations to conference papers and journal articles is in relationship with Scopus citations?
4. How long it takes for patents to cite conference papers compared to journal articles?

Method

In response to research questions, 111,801 sample of records indexed in Scopus, including 59,693 (53%) journal articles and 52,108 (47%) conference papers published in two distinct years of 1996 and 2006, were exported across 30 disciplines by August 2016. The sample included about 1000 records for each year and each document type when possible (Table 2), though due to integrating data with pilot investigations there are more papers tested for some fields such as in industrial engineering. Scopus was used to collect proceeding records, due to its broader coverage of peer-reviewed conference papers compared with Web of Science (Meho & Yang, 2007). In addition, the chosen publication years are both distant enough to accrue patent citations and can make appropriate cases to compare significant changes of each document type over time (see more detailed trends in patent uptake of Scopus indexed journal articles in Kousha and Thelwall, 2017).

So far, variety of limitations were associated with gathering non-patent citations for a large-scale analysis, such as inconsistent formulation of references in patents’ content and variety of term specification for journal titles (see more in Kousha & Thelwall, 2017). Therefore, based on the acceptable accuracy level of results acquired through Bing’s Google Patent searches for collecting patent citations (Kousha & Thelwall, 2017), Google Patents (google.com/patents) was used to gather citations information. For this purpose, bibliographic information of papers was submitted to Bing search engine, formulated in queries as below, including first authors surname, three to ten words in title of work and publication year, and Google Patents website’s URL:

Mandal "Image indexing using moments and wavelets" 1996 site:google.com/patents

Due to specifying source titles in various forms in patent reference, journal and conference titles are not included in the search queries. At the time, September 2016, the data was collected via Webometric Analyst software program (lexiurl.wlv.ac.uk) submitting queries to Bing Search
Identifying Patent Families. Google Patents (GP) claims to cover patents from six major offices (google.com/patents); therefore, its citation results tend to find patents from across variety of offices although majority of results represented US patents (Kousha & Thelwall, 2017). Results also sometimes include patents with repeated content from one patent family (for example two patents of www.google.com/patents/US6622470 and www.google.com/patents/US6910335 in Bing search results for one articles had the same titles and priority dates). Raw patent counts can provide an indicator of technological productivity, however, once many patents are extensions of a similar patent family, using the raw counts is criticized for over representing inventions (Trajtenberg, 1990). Some studies have attempted to avoid the problem of over-counting patents by calculating patent family normalized counts of inventions in company and country levels (Dernis & Khan, 2004; Park & Hingley, 2009), particularly in fields with large patent families such as in pharmaceutics (Messinis, 2011). Therefore, in order to normalize Google Patent citation counts to papers in current research, “Priority date” of citing patents was recognised because a “basic patent family” claims an equal priority date across all patent extensions and equivalents in different countries (“Derwent World Patents Index patent family,” 2017). To do this, firstly, all Bing’s recognized Google Patent pages were crawled and saved using SocSciBot crawler software (socscibot.wlv.ac.uk). Then, using the service of “Extract SocSciBot crawl info” in Webometric Analyst the details of patents was extracted from saved web pages including priority date of patents to recognize patent families and normalize GP counts based on unique families referencing them. In response to forth research question, priority date as well as filing and publication dates of patents were also identified to investigate how long it takes for papers to appear as patent references.

Limitations
One limitation of this method is that titles of conference papers could interfere with non-patent references such as journals articles with similar titles and author. Bar-Ilan (2010) and Michels and Fu (2014) pointed out that publication of conference papers in periodicals lead to problem of double counting papers. Authors of conference papers that have provided a primary source of information usually intend to revise and publish in peer-reviewed periodicals (Drott, 1995), such as about 33% of proceedings at International Society for Informetrics and Scientometrics (Aleixandre-Benavent, González-Alcaide, Miguel-Dasit, Navarro-Molina, & Valderrama-Zurián, 2009). This seems to be important since about one-fourth (approximately 2 million in 7.7 million) of Scopus indexed conference proceedings records are published in special issues of periodicals (Scopus Content Coverage Guide, 2016) which could have influenced results in current research, because source title could not be included in search query (more above). Another issue was the overlap of paper titles with similar patent titles. These patents had older priority date (over 4 years older) than publication date of paper, so they could be recognised and after checking discarded as false results (6.5%). Another limitation is that only two years are checked and results can be influenced by particular events on these years. In addition to this, the differences between documents types in each year were rather inconsistent which can be influenced by changes in topics and focuses of conference events.
Findings

The Google Patents’ searches identified 6,571 patents which referenced overall 2,722 (2.5%) papers, including 1,703 (2.9%) journal articles and 988 (1.9%) conference papers. Based on the priority date of retrieved patents 3,070 (60%) and 2,033 (40%) patent families were recognized, which referenced journal articles and conference papers, respectively. Table 1 demonstrates GP uptake of papers across 30 Scopus subjects for 1996 and 2006 together. The least proportion of journal articles with GP citations was 0.4% (5 papers) in Automotive Engineering and the highest was 12.6% (246) in Biomedical Engineering, while the second highest were just below half this proportion (5.4%) in both Biotechnology (107) and Computer Vision and Pattern Recognition (108). In terms of conference papers, this proportions ranged between 0.3% in three fields of Artificial Intelligence (3), Environmental Engineering (4), and Ceramics and Composites (3) and 4.5% in Computer Vision and Pattern Recognition (91). Table 2 makes comparable the means of GP and Scopus citations across fields. Regarding the whole samples, medians of GP citations were all 0, while the mean of citations ranged between 0.01 and 0.48. In fourteen fields, the difference of average patent uptake of conference papers and journal articles were insignificant, with the highest close averages of GP citations (including zero counts) occurring in Computer Vision and Pattern Recognition (0.10 journal article and .11 conference papers) and the lowest close means in Building and Construction (both .01). Consistent with previous findings (Kousha & Thelwall, 2017), Biomedical Engineering had the highest mean patent uptake of journal articles (0.48), which is found significantly more than conference papers’ average patent uptake (0.06) (Mann-Whitney test significant at p < .001). The second most significant difference was in Computer Graphics and Computer-Aided Design, where journal articles received significantly more GP family normalised counts (0.15, p < .001) than conference papers (0.04). Similarly, in 14 other fields average patent uptake of documents was significantly different (Table 2), while in majority of them (9 fields) patent uptake of journal articles were significantly higher than conference papers. In the remaining five fields conference papers advantaged significantly more GP citations, among which the highest significant differences were apparent in Geotechnical Engineering and Engineering Geology (mean 0.06 conference papers vs. 0.01 articles, Mann-Whitney test significant at p < .001), Automotive Engineering (0.05 vs. 0, p < .001), and Computer Science Applications (0.07 vs. 0.03, p < .01). Furthermore, in most fields differences between patent uptake of document types have been erratic over the years, changing superiority of each document type from 1996 to 2006, however, most significant differences were apparent only in 1996.

In addition to testing difference of means using Mann Whitney test (more above) a regression model was also used to assess whether document type is predictive of patent uptake. For citation data, the ordinary least squares regression model was used as it is recommended on transformed data through natural logarithm of citations +1 (Thelwall, 2016) instead of zero-inflated negative binomial regression model, due to capability to return false influences (Thelwall & Wilson, 2014). After regressing transformed data by using ordinary least squares most significant difference shown in Mann Whitney was repeated with almost similar significance level (Table 2); however, in six fields there were variations. In addition to above results, conference papers in Civil and Structural Engineering as well as Industrial Engineering seemed to advantage significantly more GP citations, whereas journal articles in Signal Processing benefited from significantly more patent uptake (all significant at p < .05). In contrast, Information Systems, Safety, Risk, Reliability and Quality, and Atomic and Molecular Physics, and Optics did not
show significant differences in regression analysis. In terms of Scopus citations, however, consistent with finding from Web of Science’s Science Citation Index (Lisée et al., 2008) there was significantly fewer citations to conference papers than to journal articles across all disciplines ($p < 0.001$).

Table 1. Papers with at least one citations in Bing’s Google Patent searches and Scopus across document types and sample sizes – Above lines indicate journal articles (ar) and below conference papers (cp).

<table>
<thead>
<tr>
<th>Scopus Subject based on ASJC</th>
<th>Papers with GP citations (% cited) citing patent families: raw GP counts 1996+2006</th>
<th>Papers with Scopus citations (% cited) citing papers 1996+2006</th>
<th>Total sample size (actual % of ar &amp; cp$^a$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biochemistry</td>
<td>33 (2.6%) 47:52</td>
<td>1187 (94%) 34,189</td>
<td>1,263 (97%) 1,954 (3%)</td>
</tr>
<tr>
<td>Biophysics</td>
<td>80 (4%) 110:128</td>
<td>1966 (93%) 55,842</td>
<td>1,999 (96%) 803 (4%)</td>
</tr>
<tr>
<td>Biotechnology</td>
<td>33 (4.1%) 50:54</td>
<td>1824 (91%) 52,881</td>
<td>2,000 (89%) 1,118 (11%)</td>
</tr>
<tr>
<td>Bioengineering</td>
<td>106 (5.3%) 169:201</td>
<td>1815 (91%) 49,182</td>
<td>2,000 (63%) 1,983 (37%)</td>
</tr>
<tr>
<td>Artificial Intelligence</td>
<td>48 (2.4%) 93:126</td>
<td>1660 (83%) 59,905</td>
<td>2,000 (51.5%) 1,005 (48.5%)</td>
</tr>
<tr>
<td>Computer Graphics and</td>
<td>95 (4.8%) 250:307</td>
<td>1642 (82%) 49,223</td>
<td>1,996 (61%) 1,022 (39%)</td>
</tr>
<tr>
<td>Computer-Aided Design</td>
<td>16 (1.6%) 37:38</td>
<td>519 (37%) 6,349</td>
<td>1,577 (45%) 1,519 (55%)</td>
</tr>
<tr>
<td>Computer Networks and</td>
<td>37 (2.3%) 72:100</td>
<td>1212 (77%) 26,224</td>
<td>1,994 (46%) 1,986 (36%)</td>
</tr>
<tr>
<td>Communications</td>
<td>12 (0.8%) 37:54</td>
<td>556 (39%) 5,066</td>
<td>1,577 (45%) 1,519 (55%)</td>
</tr>
<tr>
<td>Computer Science Applications</td>
<td>28 (1.4%) 54:67</td>
<td>1694 (85%) 45,500</td>
<td>2,000 (46%) 2,016 (54%)</td>
</tr>
<tr>
<td>Computer Vision and Pattern Recognition</td>
<td>108 (5.4%) 176:202</td>
<td>1534 (77%) 64,951</td>
<td>1,994 (46%) 2,006 (54%)</td>
</tr>
<tr>
<td>Information Systems</td>
<td>59 (3%) 105:137</td>
<td>1573 (79%) 47,354</td>
<td>2,000 (53%) 1,034 (47%)</td>
</tr>
<tr>
<td>Signal Processing</td>
<td>62 (3.1%) 127:166</td>
<td>1497 (75%) 41,070</td>
<td>1,998 (50.5%) 1,983 (49.5%)</td>
</tr>
<tr>
<td>Software</td>
<td>82 (4.1%) 183:231</td>
<td>1549 (78%) 37,729</td>
<td>1,997 (30%) 1,989 (70%)</td>
</tr>
<tr>
<td>Geotechnical Engineering and Engineering Geology</td>
<td>14 (0.7%) 16:17</td>
<td>1178 (59%) 21,551</td>
<td>1,999 (64%) 1,986 (36%)</td>
</tr>
<tr>
<td>Energy Engineering and Power Technology</td>
<td>28 (1.4%) 36:43</td>
<td>1390 (70%) 30,412</td>
<td>1,998 (62%) 1,971 (38%)</td>
</tr>
<tr>
<td>Aerospace Engineering</td>
<td>15 (0.8%) 20:22</td>
<td>1189 (60%) 17,742</td>
<td>1,995 (34.5%) 1,966 (65.5%)</td>
</tr>
<tr>
<td>Automotive Engineering</td>
<td>5 (0.4%) 5:5</td>
<td>747 (55%) 8,112</td>
<td>1,362 (20%) 1,989 (80%)</td>
</tr>
<tr>
<td>Biomedical Engineering</td>
<td>245 (12.6%) 660:1063</td>
<td>1388 (71%) 46,327</td>
<td>1,946 (78%) 2059 (22%)</td>
</tr>
<tr>
<td>Civil and Structural</td>
<td>11 (0.6%) 20:20</td>
<td>1621 (81%) 31,024</td>
<td>2,000 (60.5%) 1,884 (39.5%)</td>
</tr>
<tr>
<td>Engineering</td>
<td>19 (1%) 64:92</td>
<td>688 (50%) 4,140</td>
<td>2,000 (41%) 1,971 (59%)</td>
</tr>
<tr>
<td>Control and Systems</td>
<td>39 (2%) 91:105</td>
<td>1498 (75%) 43,167</td>
<td>2,000 (41%) 1,971 (59%)</td>
</tr>
<tr>
<td>Engineering</td>
<td>36 (1.8%) 68:100</td>
<td>993 (45%) 8,259</td>
<td>2,000 (41%) 1,971 (59%)</td>
</tr>
<tr>
<td>Electrical and Electronic Engineering</td>
<td>46 (2.3%) 78:86</td>
<td>1724 (86%) 55,508</td>
<td>1,999 (42%) 2,961 (58%)</td>
</tr>
</tbody>
</table>
Correlation between Google Patents and Scopus Citations

Correlation between Scopus and GP citations, shown in Table 2, was very low and slightly high in earlier year. The Spearman’s correlation coefficients were slightly higher in conference papers (ranging between 0.06 and 0.279) than in journal articles (between -0.081 and 0.233), despite inconsistencies. Although in one-third of fields the correlations were not significant in both document types, the highest significant correlations in conference papers were in 1996 Computer Graphics and Computer-Aided Design (r = 0.251, p < 0.01) and 2006 Biophysics (r = 0.279, p < 0.01). In terms of journal articles, the highest correlations were in 1996 Computer Vision and Pattern Recognition (r = 0.233, p < 0.01) and 2006 Bioengineering and Biotechnology (in both r = 0.218, p < 0.01).

Age of Papers by Patent Uptake

Figure 1 indicates age of papers when they were cited in patents (deviation of paper’s publication date from patent’s priority date). Priority date of patents citing about 23% of scholarly papers with Google Patents’ citations were in advance of their publication date. Furthermore, proportions of conference papers cited in patents tended to reach its maximum rate (about 0.7%) 2 years after they were published; however, only about 1 year after they formally appeared to journal articles (about 0.8%). Filing dates were not available in 9.4% (647) of Google Patent pages, though they indicated that it took almost 4 years for journal articles and 6 years for conference papers to maximise as patent references. Publication date of patents indicates that it took about 8 years for most citing patents to be published, although there was
a minor peak for both conference papers and journal articles uptake about 18 years later which obviously relate to 1996 papers and might have been influenced by the granting processes and publication periods of office(s).

Figure 1. Age of conference papers vs. journal article in terms of technological uptake

Discussions and Conclusion

RQ1. Regarding first research question, there was very low uptake of both journal articles (2.9%) and conference papers (1.9%) in patents, while the difference in average of citations (0.06 vs. 0.05 patent citations, respectively) was not significant and overall medians (both 0) and geometric means (0.02) were equal. This suggest that although patent impact constitutes a tiny proportion of research impact, conference papers does not generally appear different from journal articles if technological value of public science is meant to be assessed, and this makes an important evidence as in terms of conventional citations conference papers were significantly disadvantaged form of publication compared to journal articles (Lisée et al., 2008).

RQ2. As addressed in the second research question, the number of fields with disproportionately more patent uptake of one document type was of 30 fields only 14: nine fields with journal articles and 5 with conference papers more advantaged. The broadest difference was shown by Biomedical Engineering where journal articles received disproportionately more patent citations that conference papers. The two fields in which conference papers should be sensitively taken into account when patent impact is being assessed are Geotechnical Engineering and Engineering Geology and Automotive Engineering where journal articles was genuinely playing the least role to influence technology, at least in the years assessed. Computer Science Applications is the next field where still there is a significant patent uptake of conference papers, although journal articles also play a role.

RQ3. With regards to third research question, the relationships between GP and Scopus citations to conference papers was all weak, although slightly stronger than journal articles. The correlations were marginally higher in fields with weaker connection to engineering and more scientific nature such as Biophysics and Biochemistry, whereas consistent with prior research (Kousha & Thelwall, 2017) correlation in more engineering-related fields was very week and usually not significant, likely to suggest the broad difference between research with significantly high technological impact and scholarly influence.

RQ4. Regarding the forth question, as Figure 1 illustrates conference paper uptake in patents peaked slightly afterwards journal articles, which is slightly surprising because conference
papers tend to be published relatively sooner than journal articles. However, this evidence probably reflects that most patentable ideas published in journal articles was being processed along with patenting; whereas, the latency in patent uptake of conference papers might be evidence of freshness of ideas offered through scientific meetings. However, any implication about the patent uptake of scholarly research should be made cautiously, because references of academic papers might be made by examiner rather than inventors due mandatory aspects of involving closely related literature (Jaffe, Trajtenberg, & Fogarty, 2000) and might have less direct involvement in citing patent’s innovative idea (Callaert, Pellens, & Van Looy, 2014).

**Conclusion.** Ultimately, findings are providing the evidence that conference papers can be effectively used along with journal articles in order to assess patent uptake of academic research in most fields. However, despite this statistical evidence, another research is required to speculate how much vigorous are the patent cited ideas in conference papers compared with journal articles. It also should be considered that patent impact of conference papers could become apparent slightly later than they do for journal articles.

### Table 2. Median, mean, geometric mean and maximum of Google Patent (GP) family normalised results and Scopus citations – Spearman’s correlation coefficients in two years – Above lines indicate journal articles and below conference papers

<table>
<thead>
<tr>
<th>Scopus Subjects</th>
<th>Google Patent citations</th>
<th>Scopus citations</th>
<th>Spearman’s r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All sample: Median</td>
<td>Without 0's Median</td>
<td>1996</td>
</tr>
<tr>
<td></td>
<td>(Geometric mean) Max.</td>
<td>(Mean)</td>
<td>[Mean]</td>
</tr>
<tr>
<td>Biochemistry</td>
<td>0 (0.02) 7 [0.04]</td>
<td>1 (1.6)</td>
<td>17 (14.9) 374 [27.1]</td>
</tr>
<tr>
<td></td>
<td>0 (0.03) 14 [0.05]</td>
<td>1 (1.6)</td>
<td>10 (8.4) 862 [23.2]</td>
</tr>
<tr>
<td>Biophysics</td>
<td>0 (0.04) 6 [0.06]</td>
<td>1 (1.6)</td>
<td>16 (14) 1553 [27.9]</td>
</tr>
<tr>
<td></td>
<td>0 (0.04) 4 [0.07]</td>
<td>1 (1.6)</td>
<td>11 (9.7) 2150 [32.1]</td>
</tr>
<tr>
<td>Biotechnology</td>
<td>0 (0.05) 28 [0.10]***</td>
<td>1 (1.8)</td>
<td>15 (12.1) 765 [26.4]</td>
</tr>
<tr>
<td></td>
<td>0 (0.03) 13 [0.06]</td>
<td>1 (1.9)</td>
<td>0 (1.5) 276 [9.3]</td>
</tr>
<tr>
<td>Bioengineering</td>
<td>0 (0.05) 28 [0.10]***</td>
<td>1 (1.8)</td>
<td>11 (10.4) 858 [24.6]</td>
</tr>
<tr>
<td></td>
<td>0 (0.03) 10 [0.05]</td>
<td>1 (2.0)</td>
<td>0 (0.9) 340 [3.1]</td>
</tr>
<tr>
<td>Artificial Intelligence</td>
<td>0 (0.03) 10 [0.06]***</td>
<td>1 (2.4)</td>
<td>9 (8.2) 1775 [30]</td>
</tr>
<tr>
<td></td>
<td>0 (0) 0 [0]</td>
<td>1 (1.3)</td>
<td>0 (0.8) 248 [2.4]</td>
</tr>
<tr>
<td>Computer Graphics and Computer-Aided Design</td>
<td>0 (0.06) 23 [0.15]***</td>
<td>1 (3.1)</td>
<td>7 (7.1) 1301 [24.7]</td>
</tr>
<tr>
<td></td>
<td>0 (0.02) 16 [0.04]</td>
<td>1 (2.4)</td>
<td>1 (1.5) 423 [6.2]</td>
</tr>
<tr>
<td>Computer Networks and Communications</td>
<td>0 (0.03) 16 [0.06]***</td>
<td>1 (2.4)</td>
<td>4 (4.7) 1727 [16.6]</td>
</tr>
<tr>
<td></td>
<td>0 (0.01) 18 [0.03]</td>
<td>1.5 (3.7)</td>
<td>0 (0.8) 228 [3.3]</td>
</tr>
<tr>
<td>Computer Science Applications</td>
<td>0 (0.01) 15 [0.03]</td>
<td>1 (2.2)</td>
<td>8 (7.4) 2305 [22.8]</td>
</tr>
<tr>
<td></td>
<td>0 (0.03) 24 [0.07]***</td>
<td>1 (2.5)</td>
<td>0 (0.7) 418 [2.6]</td>
</tr>
<tr>
<td>Computer Vision and Pattern Recognition</td>
<td>0 (0.05) 18 [0.10]</td>
<td>1 (2.2)</td>
<td>7 (7.1) 2543 [32.6]</td>
</tr>
<tr>
<td></td>
<td>0 (0.05) 18 [0.11]</td>
<td>1 (2.4)</td>
<td>1 (2.2) 391 [7.2]</td>
</tr>
<tr>
<td>Information Systems</td>
<td>0 (0.03) 14 [0.07]***</td>
<td>1 (2.2)</td>
<td>5 (6) 2394 [23.7]</td>
</tr>
<tr>
<td></td>
<td>0 (0.03) 16 [0.07]</td>
<td>1 (4.1)</td>
<td>0 (0.9) 358 [5]</td>
</tr>
<tr>
<td>Signal Processing</td>
<td>0 (0.03) 29 [0.08]***</td>
<td>1 (2.5)</td>
<td>5 (5.1) 2543 [20.6]</td>
</tr>
<tr>
<td></td>
<td>0 (0.02) 14 [0.04]</td>
<td>1 (2.0)</td>
<td>1 (1.3) 426 [3.8]</td>
</tr>
<tr>
<td>Software</td>
<td>0 (0.05) 23 [0.11]***</td>
<td>1 (2.6)</td>
<td>5 (5) 2063 [18.9]</td>
</tr>
<tr>
<td></td>
<td>0 (0.03) 24 [0.09]</td>
<td>1.5 (3.3)</td>
<td>1 (1.9) 1072 [7.8]</td>
</tr>
<tr>
<td>Geotechnical Engineering and Engineering Geology</td>
<td>0 (0.01) 2 [0.01]</td>
<td>1 (1.2)</td>
<td>2 (2.9) 485 [10.8]</td>
</tr>
<tr>
<td></td>
<td>0 (0.02) 21 [0.06]***</td>
<td>1 (2.6)</td>
<td>0 (0.4) 117 [1.3]</td>
</tr>
<tr>
<td>Energy Engineering and Power Technology</td>
<td>0 (0.01) 5 [0.02]</td>
<td>1 (1.5)</td>
<td>5 (4.7) 378 [15.2]</td>
</tr>
<tr>
<td></td>
<td>0 (0.01) 5 [0.02]</td>
<td>1 (1.8)</td>
<td>0 (0.4) 174 [1.5]</td>
</tr>
<tr>
<td>Aerospace Engineering</td>
<td>0 (0.01) 4 [0.01]</td>
<td>1 (1.4)</td>
<td>2 (2.5) 554 [8.9]</td>
</tr>
<tr>
<td></td>
<td>0 (0.01) 10 [0.02]</td>
<td>2 (2.8)</td>
<td>0 (0.2) 54 [0.5]</td>
</tr>
<tr>
<td>Scopus Subjects</td>
<td>Google Patent citations</td>
<td>Scopus citations</td>
<td>Spearman’s r</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>----------------------------------------------</td>
<td>--------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td></td>
<td>All sample: Median (Geometric mean) Max. [Mean]</td>
<td>Without 0’s Median (Geometric mean) Max. [Mean]</td>
<td>1996</td>
</tr>
<tr>
<td>Automotive Engineering</td>
<td>0 (0.1) 1 [0]</td>
<td>1 (1.0)</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>0 (0.02) 15 [0.05]*** R</td>
<td>1 (2.1)</td>
<td></td>
</tr>
<tr>
<td>Biomedical Engineering</td>
<td>0 (0.17) 57 [0.48]*** R</td>
<td>2 (3.8)</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>0 (0.02) 16 [0.06]</td>
<td>1 (4.3)</td>
<td></td>
</tr>
<tr>
<td>Civil and Structural Engineering</td>
<td>0 (0.00) 9 [0.01]</td>
<td>1 (1.8)</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>0 (0.01) 16 [0.04]</td>
<td>7 (4.3)</td>
<td></td>
</tr>
<tr>
<td>Control and Systems Engineering</td>
<td>0 (0.06) 1 [0.01]</td>
<td>2 (5.6)</td>
<td>1.45**</td>
</tr>
<tr>
<td></td>
<td>0 (0.02) 11 [0.04]</td>
<td>1 (2.4)</td>
<td></td>
</tr>
<tr>
<td>Electrical and Electronic Engineering</td>
<td>0 (0.02) 13 [0.04]</td>
<td>1 (1.8)</td>
<td>0.072**</td>
</tr>
<tr>
<td></td>
<td>0 (0.02) 41 [0.05]</td>
<td>11 (9.1)</td>
<td></td>
</tr>
<tr>
<td>Industrial Engineering</td>
<td>0 (0.02) 17 [0.04]</td>
<td>1 (2.5)</td>
<td>0.71**</td>
</tr>
<tr>
<td></td>
<td>0 (0.03) 26 [0.08]R</td>
<td>2.6 (2.6)</td>
<td></td>
</tr>
<tr>
<td>Safety, Risk, Reliability and Quality</td>
<td>0 (0.01) 6 [0.03]*</td>
<td>1 (1.1)</td>
<td>0.068*</td>
</tr>
<tr>
<td>Building and Construction</td>
<td>0 (0.01) 9 [0.01]</td>
<td>1 (1.9)</td>
<td></td>
</tr>
<tr>
<td>Environmental Engineering</td>
<td>0 (0.02) 12 [0.03]*** R</td>
<td>1 (1.8)</td>
<td>1.147**</td>
</tr>
<tr>
<td></td>
<td>0 (0.04) 4 [0.01]</td>
<td>1.5 (2.0)</td>
<td></td>
</tr>
<tr>
<td>Ceramics and Composites</td>
<td>0 (0.02) 11 [0.03]*** R</td>
<td>1 (2.0)</td>
<td>0.122**</td>
</tr>
<tr>
<td></td>
<td>0 (0.00) 0 [0]</td>
<td>1 (1.3)</td>
<td></td>
</tr>
<tr>
<td>Electronic, Optical and Magnetic</td>
<td>0 (0.01) 8 [0.03]</td>
<td>1 (2.5)</td>
<td>0.092**</td>
</tr>
<tr>
<td>Materials</td>
<td>0 (0.01) 14 [0.04]</td>
<td>1 (2.4)</td>
<td></td>
</tr>
<tr>
<td>Neuroscience(all)</td>
<td>0 (0.03) 6 [0.05]</td>
<td>1 (1.6)</td>
<td>0.139**</td>
</tr>
<tr>
<td></td>
<td>0 (0.02) 8 [0.04]</td>
<td>1 (1.7)</td>
<td></td>
</tr>
<tr>
<td>Condensed Matter</td>
<td>0 (0.01) 10 [0.03]</td>
<td>1 (1.7)</td>
<td>0.124**</td>
</tr>
<tr>
<td>Physics</td>
<td>0 (0.02) 19 [0.06]</td>
<td>1 (2.8)</td>
<td></td>
</tr>
<tr>
<td>Instrumentation</td>
<td>0 (0.02) 6 [0.03]*** R</td>
<td>1 (1.7)</td>
<td>0.111**</td>
</tr>
<tr>
<td></td>
<td>0 (0.01) 10 [0.02]</td>
<td>1 (1.9)</td>
<td></td>
</tr>
<tr>
<td>Atomic and Molecular Physics, and</td>
<td>0 (0.01) 14 [0.03]</td>
<td>1 (2.9)</td>
<td>0.107**</td>
</tr>
<tr>
<td>Optics</td>
<td>0 (0.02) 5 [0.03]*</td>
<td>1 (1.7)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0 (0.02) 57 [0.06]</td>
<td>1 (2.3)</td>
<td>0.134**</td>
</tr>
<tr>
<td></td>
<td>0 (0.02) 41 [0.05]</td>
<td>7 (6.8)</td>
<td></td>
</tr>
</tbody>
</table>

* Mann Whitney test comparing means of document types (and Spearman’s correlation coefficient) significant at p < 0.05; ** p < 0.01; *** p < 0.001; R significant positive coefficient in regression model; Bold correlation coefficients shows the larger correlation within year.

Acknowledgments
The author would like to appreciate the researchers, particularly professor Michael Thelwall, at the Statistical Cybermetrics Research Group of the University of Wolverhampton for their invaluable advices on the current research.

References


doi:10.1002/asi.23821


Schmoch, U. (1993). Tracing the knowledge transfer from science to technology as reflected in patent indicators. *Scientometrics, 26*(1), 193-211.


Impact of multidisciplinary research on innovation

David Campbell 1 Brooke Struck 2 Chantale Tippett 3 Guillaume Roberge 4

1 david.campbell@science-metrix.com
Science-Metrix (Canada)

2 brooke.struck@science-metrix.com
Science-Metrix (Canada)

3 chantale.tippett@science-metrix.com
Science-Metrix (Canada)

4 guillaume.roberge@science-metrix.com
Science-Metrix (Canada)

Abstract
Governmental initiatives capitalising on multidisciplinary (or interdisciplinary) research are growing in number. They are motivated by the increasingly shared view among academics and policymakers that this mode of research, by favouring innovation in firms, will fuel the economic competitiveness of nations through job creation and increased revenues. However, there is an obvious lack of empirical evidence supporting the connection between multidisciplinary research and innovation. This study partly fills this gap by addressing the following question: Is the knowledge disclosed in a scientific publication more likely to be taken up in innovation (i.e. cited in the patent literature) as its multidisciplinarity index increases? The results thus obtained clearly show, in the aggregates, that multidisciplinarity increases the odds of research results being useful to innovation, thereby supporting existing R&I policy interventions or paving the way for new ones. However, because uptake in innovation, as measured through patent citations to scientific articles, remains a relatively rare phenomenon, multidisciplinary research is not an effective predictor of an individual article being useful to innovation; science is experimental, and the innovation outcome of an individual project cannot be guaranteed ahead of time merely through the extent of disciplinary mixing among the participating researchers.

Conference topic
Science policy and research assessment, Citation and co-citation analysis, Patent analysis, Knowledge discovery and data mining

Introduction
The past decade has seen a significant change in the organisation and management of scientific research. Although research has traditionally been organised into specialised disciplines, more and more governmental initiatives are emerging that aim to break the disciplinary silos (Van Rijnsoever & Hessels, 2011; Allmendinger, 2015). Various approaches are reported in the literature to operationalise this mode of research, the most common being interdisciplinary research (IDR) (Sonnenwald, 2007; Wagner et al., 2011).

The underlying rationale motivating IDR initiatives resides in an increasingly shared view among academics and policymakers: by uncovering innovative solutions residing outside the context in which they emerged, partnerships integrating multiple scientific cultures and bodies of knowledge help foster new lines of thought (i.e. the emergence of new disciplines) and help tackle and solve today’s increasingly complex problems (Allmendinger, 2015; Blackwell et al., 2010; Mainzer, 2011). Blackwell et al. (2010) refer to the former application of IDR as being curiosity driven, and the latter application as being outcome driven.

By helping companies to stay ahead, or at least abreast, of the most recent developments in a rapidly changing business environment, outcome-driven IDR is perceived as having the potential to boost the competitiveness of firms and the economic well-being of nations. While the number of initiatives capitalising on outcome-driven IDR to fuel innovation in firms is
increasing, there is a glaring lack of empirical evidence to support the notion that IDR is an effective mechanism to spur innovation and longer-term job creation and competitiveness in the knowledge economy. As laid out by Allmendinger (2015):

While there are plenty of data, insights and lessons on directed research programs and organized research units at universities, we have but next to no empirical evidence on how to best stage interdisciplinarity, about the added value it may produce, and what it may take universities and research organizations to effectively cross narrow disciplinary boundaries, perspectives, and interests. The ironic bottom line is that we need both more interdisciplinarity, and more organizational experiments, to advance it, and to learn more about what is conducive to it, what works and what does not.

This study aims to partly fill this gap by investigating the type of added value that may be produced by IDR. More specifically, the following question is being addressed: To what extent is IDR positively associated with innovation, focusing on outcome-driven/applied research? Because current interventions rely mostly on assumptions derived from rational—but speculative—thinking, it is still pertinent to investigate the above question in furthering our understanding of IDR and its potential outputs/outcomes—this to further assess the relevance of existing and upcoming policy interventions. With this aim, this study’s policy question was converted into the following data mining problem: the extent to which IDR is conducive to innovation is measured by investigating if, and to what extent, the knowledge disclosed in a scientific publication has greater odds of being taken up in the patent literature as its interdisciplinarity increases. Because the study focuses on a specific type of innovation—product/process innovations as disclosed in patents—the findings are not generalizable to the full innovation landscape.

Methods
Briefly, uptake in innovation is inferred by matching non-patent references (NPRs) in patents to a bibliographic database of peer-reviewed scientific literature focusing on science and engineering fields. This variable is referred to as the ‘cited in patent’ indicator throughout this paper. The scientific literature on measuring IDR has been blooming in recent years (Porter & Rafols, 2009; Rafols & Meyer, 2010; Campbell et al., 2015; Cassi, Mescheba, & de Turckheim, 2015; Calatrava Moreno, Auzinger, & Werthner, 2016). Most of these studies attempted to measure IDR using bibliometric data extracted from large bibliographic databases of peer-reviewed literature to quantify the diversity of integrated knowledge within individual research articles. This is typically done by computing the Rao-Stirling diversity index of the material cited in an article. This index integrates the number of different subfields cited, the balance between these subfields, and the intellectual distance between them (Porter & Rafols, 2009). Because one of the main eligibility criteria for researchers applying to funding programmes targeting IDR is almost always the requirement that the project team includes researchers from diverse disciplines—and because IDR measured from co-referencing of multiple disciplines in an article (as in Porter & Rafols, 2009) can result from the work of a single author—it was decided that the Rao-Stirling diversity index would be applied to papers’ contributing disciplines as revealed by the departmental affiliations of their authors, instead of their cited subfields. To differentiate this indicator from those based on cited subfields, as per a paper’s references, it is referred to as the ‘multidisciplinariness’ (MDR) throughout this paper.

Data sources
Two sources of primary data are necessary to match NPRs in patents to individual scientific publications to create the cited in patent indicator: a patent database and a database of peer-
reviewed scientific publications. For this study, PATSTAT was selected as the patent database, limited in this instance to a dataset of USPTO patents. The Web of Science (WoS), produced by Clarivate Analytics, was selected as the database for peer-reviewed scientific publications. The WoS also includes the addresses of all authors on a publication, enabling the computation of the MDR. Only articles published in the domains of Natural Sciences and Engineering (NSE) and Health Sciences (HS) were retained for this study (as defined by Science-Metrix’ classification, see Archambault, Caruso, & Beauchesne, 2011). This is because materials from the Social Sciences and Humanities are unlikely to be taken up in innovation as disclosed in patents. Also, although reviews might be cited in patents, they typically do not disclose original research contributions. Conference papers were not available to perform this analysis.

Cited in patent indicator

The cited in patent indicator is a binary indicator, computed at the paper level. It takes the value of 0 if the publication is not cited in the patent literature and a value of 1 otherwise. A detailed discussion of the matching procedure that was implemented in this study to link the patents’ NPRs in the USPTO to scientific articles in the WoS is beyond the scope of the present paper. For a detailed presentation of this procedure, see Campbell et al. (2016).

Multidisciplinarity (MDR)

Recall that the MDR is measured with the Rao-Stirling diversity index of the disciplines represented among a paper’s author addresses in the WoS. The department names appearing in the author addresses of all papers in the WoS were harmonized to 129 distinct forms representing the disciplines used in computing the MDR. The pairwise similarity matrix between the 129 disciplines is computed using the cosine similarity between the distribution vectors, across scientific subfields as per Science-Metrix’ journal-based classification, of any two disciplines. The distribution of a given discipline across subfields is obtained by counting the papers, in each subfield, that include the corresponding discipline among their author addresses. The index varies from 0 for monodisciplinary papers to 1 for highly multidisciplinary papers. Because it was not feasible to assign a cleaned department name to all author addresses in the WoS, potential biases might arise from this limitation in the data. Although the characterisation of this indicator revealed that robust inference can be performed even in cases where not all addresses have been classified (for more details, see Campbell et al., 2016), the analyses were performed using only those articles that had all their addresses classified by department. This was done to reduce the size of the dataset, and thus allow the computation of the analysis to run (see below).

Analyses

The question to be addressed in this paper can be re-formulated as follows: Is the knowledge disclosed in a scientific publication more likely to be taken up (i.e. cited) in the patent literature as its MDR increases? To address this question, it was decided that a logistic regression would be appropriate for the type of data we are dealing with: logistic regression supports multiple discrete and/or continuous predictors (MDR is a continuous predictor and additional controls are to be added to the model) and is suitable for binary outcomes (a paper being or not being cited in patents).

Three additional variables were created: the subfield of an article, the number of contributing authors, and the number of contributing countries on an article. These were inserted in the model to control for other networking effects that could supplant the effect associated with the

1 http://www.science-metrix.com/en/classification
crossing of disciplinary boundaries in research teams. For example, an increased citation rate in patents could be due simply to the increase in the number of researchers or countries involved in a research project, rather than to the diversity of disciplines represented among them. Another factor that likely influences the citation in patent outcome is related to inter-sectoral cooperation. For example, is the publication resulting from a public–private partnership? At the time of submitting this paper, this variable had not yet been prepared and included in the study’s model. Further, due to space limitations in the current paper, the additional materials related to this analysis will be added to the oral presentation.

In the context of this study, it is important to note that the event to be modelled—the citation of an article in patents—is rare. For instance, only 2% of 2008 articles in the NSE and HS, as indexed in the WoS, have been cited in USPTO patents (see Table 3 in Results section). With dozens of times fewer 1s than 0s (i.e. cited in patents vs. not cited in patents), most logistic regression tools would likely underestimate the probability of citation in patents even in the presence of sample sizes in the thousands (King & Zeng, 2001). An algorithm that provides an accurate estimate of the binary logistic regression’s coefficients in such circumstances is Firth’s biased reduced logistic regression using penalised likelihood (Firth, 1993). The R implementation of this algorithm was selected for this study (see ‘logistf’ package created by Heinze and colleagues, 2015).

Since the algorithm used to estimate the parameters of the constructed regression model is computationally intensive and could not run on the entire dataset, the dataset was reduced using three consecutive steps. First, the analysis was restricted to the most recent publication year of scientific articles (in the WoS) for which a sufficiently long time window was available to capture their citations in US patents (in the USPTO); this was to ensure that the analysis would reveal contemporary effects. Since it can take, on average, about 3.5 years for patents to be granted from their application date at the USPTO, and because relevant references to scientific articles can be added by patent examiners in this period, the time window over which citations can be captured should minimally extend over a five-year period—that is, four years to account for the granting process at the USPTO, and one year to account for the fact that papers can be published late in a given year. As complete data on patents were available up to 2014, this means that the citation windows of articles published after 2010 (citation window of five years; 2010 through 2014) would be too short. To balance data accuracy with the need to focus on contemporary effects, it was decided that a minimum window of seven years would be used, restraining the analysis to 2008 articles. Articles published in earlier years could have also been used, but this would have necessitated including an additional control variable to the model (i.e. publication year) to account for the fact that older papers have had the chance to accumulate citations over a longer period. Since the model was already taking a long time to run, the focus was placed on 2008 articles only. This led to a dataset of 402,916 articles times five variables (subfield, MDR, number of authors, number of countries and citation in patents) for a total of about 2 million data points. Following this restriction, the dataset was still too large to run the selected analysis method.

As a second step, limiting the analysis to only those articles with all their addresses classified by discipline (or department) reduced the size of the dataset by an additional 37% (from 402,916 to 255,372 articles). Filtering using such a criterion also carries the benefit of eliminating a potential source of biases in the measurement of multidisciplinarity at the individual paper level. Conversely, such a filter may bias the population towards papers with fewer authors, and hence towards papers that are less multidisciplinary; the more addresses there are on a paper, the more difficult it is to clean all addresses. To assess the robustness of the conclusions drawn in this study, the analyses presented in this paper could be run again using a random sample of papers that have at least 5 addresses with a cleaned department or that have a cleaned department assigned to at least 50% of their addresses when they have fewer than 10 addresses. That said,
since the analysis was possibly focused on a portion of the population of papers that is less multidisciplinary, it is anticipated that the results presented here are conservative.

With the above filter applied to the dataset, the computer (4 cores/8 threads Intel Xeon CPU E3-1240 v2 processor (3.40GHz) running a 64-bit version of Windows Server 2008 R2 with 32GB of RAM) crashed after a day of computing. Accordingly, as a third step, this restricted dataset was randomly downsampling to 100,000 articles so the analysis could run successfully.

**Global analysis:** In performing the analysis for all 132 NSE and HS subfields combined, the subfield of scientific articles was added to the model as a control (dummy) variable. This was important to control for differences that prevail in the citation practices across technological fields (e.g., some fields generally include more references to the scientific literature than other fields do). As the global model was statistically significant ($p$-value < 0.05), the model was then run again for each subfield separately.

**Analysis by subfield:** Out of the 132 NSE and HS subfields, 44 (one third) that had at least 30 papers cited in patents were retained for analysis. The subfields were filtered since a small number of cases on the rarer of the two outcomes in binary logistic regression can lead to underestimation of the odds ratios. These 44 subfields accounted for 65% of all papers in the combined subfields. All articles with 100% of their addresses classified by discipline (or department) were kept in the analysis (i.e., 164,972 articles out of 255,372; the data were not downsampled). Only the subfields with a statistically significant relationship between the multidisciplinarity of peer-reviewed scientific articles and the event of being cited in patents are reported in the results section.

**Results**

*Global analysis*

When all 132 NSE and HS subfields were combined, the global model was statistically significant ($p$-value < 0.05); its associated parameters are shown in full in Table 1. Note that the lower and upper odds ratios provide the 95% confidence interval of the odds ratio for a given predictor. When the confidence interval does not overlap with 1, it means that the odds ratio is significant ($p$-value of the odds ratio < 0.05).

The odds ratio of a predictor typically shows the multiplicative factor by which the odds of citation (of an article, by a patent) increase with each full unit change in the predictor. For multidisciplinarity only, the odds ratio was re-scaled for a magnitude of change of 0.1 unit since a full unit change for MDR is highly unlikely; in contrast to the number of authors and countries, the MDR score of a paper is bounded between 0 and 1. Also, a variation of 0.1 is commonly observed across all papers; 42% of articles have a deviation from the median MDR score that is greater than or equal to 0.1. For the number of authors and countries, the odds ratio is reported for a 1-unit change, which is more commonly observed for the former relative to a 0.1-unit change for MDR (78% of papers deviate by one or more units from the median), but less commonly observed for the latter (27% of papers). How to interpret the MDR score, to digest what a 0.1-unit increase means in a practical context, is explored at length in Campbell et al. (2016).

As detailed in Table 1, the MDR of an article has a statistically significant effect on the likelihood that its results will be taken up in innovation; for instance, when the MDR score of an article increases by 0.1 unit, its likelihood of being cited in the patent literature increases by 12% (odds ratio of 1.12).
Table 1. Relationship between the multidisciplinarity, as well as the number of authors and countries, of scientific articles (2008) and their citation in the patent literature (patents issued in 2008–2014).

<table>
<thead>
<tr>
<th>Model variables (Subfield as a dummy)</th>
<th>Odds Ratio</th>
<th>Lower Odds Ratio</th>
<th>Upper Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multidisciplinarity ǂ</td>
<td>1.12</td>
<td>1.09</td>
<td>1.16</td>
</tr>
<tr>
<td>Number of authors</td>
<td>1.09</td>
<td>1.07</td>
<td>1.11</td>
</tr>
<tr>
<td>Number of countries</td>
<td>0.92</td>
<td>0.84</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: ǂ The odds ratio was re-scaled for a magnitude of change of 0.1 unit instead of 1 unit (recall that the MDR can only take values from 0 to 1).

This analysis of odds ratios was also conducted on the number of authors listed for a paper and the number of countries participating in the collaboration, to determine whether either of these features exert an effect that could suppress the influence of multidisciplinarity on the uptake of scientific knowledge in innovation. The meaning of such a change is simple to interpret for these predictors: Does having one more author, or one more participating country, involved in publishing a paper increase its likelihood of being cited in the patent literature?

The odds ratio for having an additional author is statistically significant and equals 1.09, meaning that having one more author on a paper increases the chances of citation in the patent literature by 9% (Table 1). Thus, a larger number of collaborating authors does have a positive effect on the uptake of scientific knowledge in innovation and this effect does not override the effect linked to the multidisciplinarity of a research team. International collaboration does not appear to exert a statistically significant effect on the uptake of scientific knowledge in patents.

**Analysis by subfield**

Now that a positive and statistically significant effect of MDR and the number of collaborating authors on the likelihood of being cited in the patent literature has been detected at the global level, it becomes relevant to assess if this effect holds equally well across subfields or if there are any meaningful disparities. The results are only presented for the 44 subfields with at least 30 papers cited in patents; as previously noted, subfields were filtered in this way since a small number of cases on the rarer of the two outcomes in binary logistic regression can lead to an underestimation of the odds ratios.

A summary of results at the subfield level is presented in Table 2. Of the 44 retained subfields, the model was statistically significant in 29 cases (66%). Of those 29 subfields, the odds ratio for MDR scores is statistically significant in 16 subfields (55%) and the effect is always positive in these cases. For the number of authors on a paper, the odds ratio is statistically significant in 22 of the 29 subfields (76%) for which the model is significant and the effect is positive in 95% of these cases. As noted above, international collaboration does not have a statistically significant connection with innovation, a finding that is borne out here at the subfield level: the odds ratio for number of countries is statistically significant in 9 of the 29 subfields (31%) for which the model is significant and the effect is positive in only 44% of these cases.

Focusing on only those subfields where the connections to innovation (positive or negative) are statistically significant for each predictor taken separately, we can once again look at the effects of MDR, number of authors and number of countries. On average across statistically significant subfields for MDR, an increase of 0.1 in MDR score is associated with a 32% increase in the likelihood of being cited in the patent literature (avg. odds ratio of 1.32, Table 2). A paper with one more author (a 1-unit change for number of authors) is associated, on average, with a 13% increase in chance of being cited by a patent. The inclusion of an additional country in
International collaboration is associated, on average, with a 14% increase in the chance of being cited in the patent literature. Although this is similar to the score for the number of authors, the range of the statistically significant odds ratios for the number of countries is broader, with stronger scores on the negative side effect (i.e. odds ratio below 1). The highest (significant) odds ratio for the number of countries is 1.72, but the lowest is 0.62. These values are more extreme than those for MDR (1.13 to 1.54) and the number of authors (0.92 to 1.13), meaning that there is much stronger variation from one subfield to the next in terms of the connection between international collaboration and citation in the patent literature. This finding is unsurprising given that findings presented above suggested that the connection was much weaker between international collaboration and patent citation (in the aggregate and across subfields). Only for multidisciplinarity are the effects consistently on the positive side across subfields where the predictor has a statistically significant effect—that is, multidisciplinarity never appears to lead to important decreases in the likelihood of citation in patents.

Table 2. Comparative analysis of the magnitude of effect of the multidisciplinarity, number of authors and number of countries of scientific articles (2008) on their odds of being cited in the patent literature (patents issued in 2008–2014).

<table>
<thead>
<tr>
<th>Percentage of subfields with a significant odds ratio</th>
<th>Multidisciplinarity†</th>
<th>Number of authors</th>
<th>Number of countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of subfields with an odds ratio greater than 1 (i.e. positive effects) among those with a significant odds ratio</td>
<td>100%</td>
<td>95%</td>
<td>44%</td>
</tr>
<tr>
<td>Average odds ratio across subfields where the odds ratios are significant</td>
<td>1.32</td>
<td>1.13</td>
<td>1.14</td>
</tr>
<tr>
<td>Minimum odds ratio across subfields where the odds ratios are significant</td>
<td>1.13</td>
<td>0.92</td>
<td>0.62</td>
</tr>
<tr>
<td>Maximum odds ratio across subfields where the odds ratios are significant</td>
<td>1.54</td>
<td>1.45</td>
<td>1.78</td>
</tr>
</tbody>
</table>

The results concern 66% (29) of the 44 retained subfields with a statistically significant model

| Note: † The odds ratios were re-scaled for a magnitude of change of 0.1 unit. |

In summary, multidisciplinarity is most consistently connected to patent citation, compared to the other two predictors. This finding holds both in the aggregate and at the subfield level. However, the number of authors has a significant effect in a larger set of subfields with an effect which is nearly always positive. Thus, both multidisciplinarity and the number of authors are significant contributing factor to the uptake of articles in patents. For the 16 subfields in which the multidisciplinarity–innovation link is statistically significant, the odds ratios range from a 13% increase to a 54% increase in the probability of being cited in patents for a 0.1-unit increase in MDR score. In-depth results for each of these subfields are presented in Table 3. More specifically, this table provides information on the number of cited and uncited articles in each subfield (along with the percentage of cited articles), their respective odds ratios for a 0.1-unit increase in MDR score, the upper and lower bounds of those odds ratios, and the share of articles deviating by at least 0.1 unit from the median MDR score in the given subfield. The coefficients of other variables included in the model are not shown here.
Table 3. Relationship between the multidisciplinarity of scientific articles (2008) and their citation in the patent literature (patents issued in 2008–2014).

<table>
<thead>
<tr>
<th>Subfield</th>
<th>N Uncited</th>
<th>Cited</th>
<th>Percent Cited</th>
<th>Model (p-value)</th>
<th>OR X→X+0.1 Lower</th>
<th>OR X→X+0.1 Upper</th>
<th>Share of articles with a dev. from the median multidisc. ≥ 0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>All 132 NSE &amp; HS subfields</td>
<td>98 013</td>
<td>1 987</td>
<td>2.0%</td>
<td>0.000</td>
<td>1.12</td>
<td>1.09</td>
<td>1.16</td>
</tr>
</tbody>
</table>

Global model (subfield as dummy; the odds ratio [OR] is only shown for multidisciplinarity)

<table>
<thead>
<tr>
<th>Subfield</th>
<th>N</th>
<th>Cited</th>
<th>Percent Cited</th>
<th>Model (p-value)</th>
<th>OR X→X+0.1 Lower</th>
<th>OR X→X+0.1 Upper</th>
<th>Share of articles with a dev. from the median multidisc. ≥ 0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthopaedics</td>
<td>1 859</td>
<td>34</td>
<td>1.8%</td>
<td>0.001</td>
<td>1.54</td>
<td>1.21</td>
<td>1.99</td>
</tr>
<tr>
<td>Pathology</td>
<td>1 803</td>
<td>45</td>
<td>2.4%</td>
<td>0.001</td>
<td>1.44</td>
<td>1.13</td>
<td>1.87</td>
</tr>
<tr>
<td>Nuclear Medicine &amp; Medical Imaging</td>
<td>4 231</td>
<td>106</td>
<td>2.4%</td>
<td>0.000</td>
<td>1.40</td>
<td>1.21</td>
<td>1.62</td>
</tr>
<tr>
<td>Gastroenterology &amp; Hepatology</td>
<td>3 507</td>
<td>50</td>
<td>1.4%</td>
<td>0.01</td>
<td>1.36</td>
<td>1.08</td>
<td>1.74</td>
</tr>
<tr>
<td>General Chemistry</td>
<td>3 354</td>
<td>119</td>
<td>3.4%</td>
<td>0.000</td>
<td>1.36</td>
<td>1.21</td>
<td>1.54</td>
</tr>
<tr>
<td>Ophthalmology &amp; Optometry</td>
<td>1 825</td>
<td>41</td>
<td>2.2%</td>
<td>0.023</td>
<td>1.28</td>
<td>1.03</td>
<td>1.61</td>
</tr>
<tr>
<td>Polymers</td>
<td>3 258</td>
<td>108</td>
<td>3.2%</td>
<td>0.000</td>
<td>1.28</td>
<td>1.11</td>
<td>1.48</td>
</tr>
<tr>
<td>Organic Chemistry</td>
<td>4 580</td>
<td>167</td>
<td>3.5%</td>
<td>0.000</td>
<td>1.22</td>
<td>1.12</td>
<td>1.34</td>
</tr>
<tr>
<td>Biomedical Engineering</td>
<td>1 980</td>
<td>106</td>
<td>5.1%</td>
<td>0.005</td>
<td>1.19</td>
<td>1.04</td>
<td>1.38</td>
</tr>
<tr>
<td>Neurology &amp; Neurosurgery</td>
<td>11 441</td>
<td>206</td>
<td>1.8%</td>
<td>0.000</td>
<td>1.19</td>
<td>1.07</td>
<td>1.32</td>
</tr>
<tr>
<td>Applied Physics</td>
<td>8 466</td>
<td>265</td>
<td>3.0%</td>
<td>0.000</td>
<td>1.17</td>
<td>1.08</td>
<td>1.28</td>
</tr>
<tr>
<td>Analytical Chemistry</td>
<td>3 733</td>
<td>109</td>
<td>2.8%</td>
<td>0.030</td>
<td>1.17</td>
<td>1.04</td>
<td>1.32</td>
</tr>
<tr>
<td>Oncology &amp; Carcinogenesis</td>
<td>9 026</td>
<td>350</td>
<td>3.7%</td>
<td>0.000</td>
<td>1.16</td>
<td>1.06</td>
<td>1.26</td>
</tr>
<tr>
<td>Nanoscience &amp; Nanotechnology</td>
<td>2 467</td>
<td>243</td>
<td>9.0%</td>
<td>0.000</td>
<td>1.15</td>
<td>1.05</td>
<td>1.26</td>
</tr>
<tr>
<td>Pharmacology &amp; Pharmacy</td>
<td>4 041</td>
<td>118</td>
<td>2.8%</td>
<td>0.005</td>
<td>1.14</td>
<td>1.02</td>
<td>1.29</td>
</tr>
<tr>
<td>Cardiovascular System &amp; Hematology</td>
<td>5 825</td>
<td>118</td>
<td>2.0%</td>
<td>0.001</td>
<td>1.13</td>
<td>1.00</td>
<td>1.30</td>
</tr>
</tbody>
</table>

Note: OR_{X→X+0.1} = odds ratio re-scaled for a change of 0.1 unit in the MDR score of papers.

Out of the 16 subfields where a positive and statistically significant connection exists between multidisciplinarity and the uptake of scientific knowledge in innovation, 10 (63%) are related to the Health Sciences (Table 3), where the extent of multidisciplinarity research is more pronounced (data not shown). In Table 3, the odds ratio varies from a low of 1.13 in Cardiovascular System & Hematology to a high of 1.54 in Orthopaedics. In Orthopaedics, this means that a 0.1-unit change in MDR translates into a 54% increase in the odds of an article being cited in patents. Of all the articles in this subfield, 46% deviate from the median MDR score by 0.1 unit or more. This indicates that some papers stand out by a sufficient margin in terms of MDR, significantly increasing the likelihood of their results being taken up in patents. The subfields with the most disparate spreads of MDR scores are Organic Chemistry, Biomedical Engineering, and General Chemistry, each of which has at least 50% of its articles with an MDR score 0.1 unit or more away from the median. In turn, it is reasonable to assume that a policy aiming to promote multidisciplinarity as a way to catalyse innovation could lead to a sufficient change in the MDR of supported articles in these subfields to achieve a noticeable increase in the likelihood of their findings being taken up in innovation.

Until now, the odds ratios have only been analysed for 0.1-unit changes in MDR. Although the odds ratio when moving from an MDR of 0.1 to 0.2 is the same as when moving from an MDR of 0.5 to 0.6, it is worth noting that the relationship between an odds ratio and the magnitude of change (e.g. 0.1 unit, 0.2 unit, …, 1 unit) in the predictor (i.e. MDR) is exponential, making it relevant to investigate the magnitude of an effect that can be achieved when doubling, tripling or even quadrupling the MDR of a paper. In doing so, the analysis focused specifically on the subfield of Orthopaedics, which has the largest odds ratio—1.54 for a 0.1-unit change in MDR. Figure 1 shows the change in odds ratio as a function of the magnitude of change in the MDR score of a 2008 article in Orthopaedics. Articles with an MDR score of 0.4 are slightly more than five times as likely to be cited in the patent literature compared to monodisciplinary papers (MDR = 0) in Orthopaedics (odds ratio = 5.57); the magnitude of this effect is 3.6 times larger than for an MDR of 0.1 (odds ratio = 1.54). Taking any baseline, an increase of 0.4 of a unit in the MDR score of an article (e.g. from 0 to 0.4, 0.1 to 0.5, 0.2 to 0.6) will lead to a 5.57 times increase in the odds that the knowledge it contains will be taken up in innovation. This is a non-negligible effect when we note that close to a fifth of 2008 articles in Orthopaedics have an
MDR score of at least 0.4, while the mode of the distribution is at MDR = 0 (see inner chart in Figure 1). For an MDR score of 0.67, the highest score observed in Orthopaedics in 2008, the odds of being cited in patents are nearly 18 times larger than for monodisciplinary papers.

**Figure 1.** Relationship between the increased probability (odds ratio) of an article being cited in the patent literature and a given change in its multidisciplinarity score (for 2008 articles in Orthopaedics, cited in patents issued in 2008–2014).

The preceding analyses of MDR’s effect on the likelihood of uptake in the patent literature, for the subfield of Orthopaedics, has shown that the relationship is exponential, with the likelihood of being cited in a patent growing more and more rapidly as the magnitude of change in MDR scores increases. But an important piece of context is still absent: the relative increases in likelihood of uptake in patents has been discussed, but the likelihoods themselves, as a function of MDR scores, have so far not been quantified. Let us turn to that point now. Because citation by a patent is a relatively rare event, individual articles needed to be sorted into suitably large bins to enable the signal to emerge clearly from the background noise. For the present analysis, Orthopaedics articles were sorted into three bins in ascending order of their MDR scores: one for the ~22% of articles with an MDR score of 0, and then two equally sized bins of ~39% each, each accounting for exactly half of the remaining articles. More bins could not be produced without reducing the number of cited articles found in each bin, which would introduce unwanted noise into the analysis.

The average multidisciplinarity of articles, as well as the frequency of cited articles, was computed for each bin, producing the data points shown in Figure 2. In this figure, the results are showing the actual odds of being cited, not the odds ratio (which quantifies the relative change in odds for a given change in MDR, but not the odds themselves). As detailed in Figure 2, the observed frequency of citation in patents for the three bins prove to be a relatively strong fit to the expected odds of being cited in patents as predicted by the logistic regression model (see the exponential prediction line in the graph), suggesting the overall adequacy of the model. However, because of the rarity of article references in patents, the predictive power of this
model for any individual article will be very low; instead, the model is better suited to predicting overall patterns in the aggregate.

Figure 2. Relationship between the increased probability (odds ratio) of an article being cited in the patent literature and a given change in its multidisciplinarity score (for 2008 articles in Orthopaedics, cited in patents issued in 2008–2014).

Note: The model was fitted using three explanatory variables: multidisciplinarity, number of authors and number of countries. In this graph, the number of authors and countries were held constant using the average number of authors (4.94) and countries (1.17) across all Orthopaedics papers. This is acceptable given the small variability observed in these quantities across the three bins (no. of authors ranges from 4.47 and 5.19; no. of countries ranges from 1.17 to 1.20) and given the non-significance of the odds ratio for these two variables. The effect of these variables is here embedded in the model intercept.

Comparing Figure 1 to Figure 2 can provide a valuable perspective on the effects of MDR on uptake in innovation. For a paper with an MDR score of 0, the likelihood of being cited by a patent is 0.5%—a 1 in 200 chance. Increasing the MDR score to 0.4, which is a significant increase in multidisciplinarity, increases the likelihood more than fivefold. However, even a fivefold increase of such a small value results in a still relatively small value: a paper with an MDR score of 0.4 has about a 2.7% chance of being cited by a patent, still only a little better than 1 in 40.

Accordingly, while it is appropriate to conclude that multidisciplinarity increases the likelihood of research being taken up in innovation, its effect is not large enough to predict with certainty that a given piece of work will contribute to new innovations (i.e. be cited in patents) solely on the basis of the disciplinary diversity of the research team. Far more factors are likely to influence such an outcome, and these would have to be introduced into the model to precisely predict the innovation outcome of an individual research article (or project). Still, the relationship that has been uncovered in this study suggests that, in some scientific (sub-)fields, promoting multidisciplinary research can significantly increase the odds of fuelling new innovations, thereby paving the way for evidence-based R&I policy intervention. Additionally, it should be noted that a rather short citation window has been used, which likely underestimates the longer-term odds of a patent citation. Consequently, policy interventions promoting multidisciplinary research may have more leverage on innovation than is suggested here.
Discussion

The present study has assessed the connection between research that crosses disciplinary boundaries and contributions to innovation through the measurement of multidisciplinarity (assessing the number, balance and diversity of disciplines or departments integrated into a research team) and the resulting chances of articles published by these research teams being taken up in the patent literature, as tracked through citations. The conclusions reached are that multidisciplinarity certainly does contribute to increased chances of uptake in innovation; this finding holds across the aggregated subfields of the NSE and HS, as well as in about 1 in 3 subfields individually. Multidisciplinarity seems to have a more robust connection to innovation than the number of countries (for international collaboration). The number of authors also appears to contribute positively to citation in patents.

Even so, this study has shown that MDR is only valuable as a predictor of general tendencies within a set of cases and cannot effectively predict whether an individual article will or will not be cited by a patent. If one assumes that uptake in innovation is not the result of a chancy system, then the factors that most strongly determine uptake for individual articles have yet to be identified. Further study would be needed to fill out this picture and hopefully identify those strongly determining elements. One such idea would be to examine the sectors in which the collaborating authors are working, seeking especially to see whether collaborations involving private-sector co-authors have a higher likelihood of providing the foundation for subsequent patents. Furthermore, it must be acknowledged that even once these determining factors are identified, the large majority of peer-reviewed publications are never cited by a patent. Uptake in innovation thus remains a relatively rare phenomenon.

Integrating these research findings back into the policy context, then, what lessons may be extracted? First, this study was initially driven by the need for some evidence to support a fundamental assumption underlying a growing number of R&I policy interventions: that crossing disciplinary boundaries helps to fertilise the ground for innovation, tilling the soil for job creation, economic growth and increased competitiveness. The evidence discovered in this study suggests that this baseline assumption is indeed borne out by facts.

It is important to note that the present study has only been a pilot test. It was conducted using a single year of peer-reviewed publications (2008) and a restricted range of years for patents (2008–2014). The cleaning of author addresses by discipline (or departments) could be improved to remove additional noise from the analysis, and the algorithm matching NPRs to scientific articles in the WoS should be improved to increase its recall while maintaining its precision. Currently, it is likely that the citation rate of scientific articles in patents is underestimated, diluting the effects measured in this project. For example, the 7-year citation window used is rather short in the context of patents, possibly underestimating the longer-term odds of a patent citation. Policy interventions promoting multidisciplinary research may thus have more leverage on innovation than is suggested in this paper. Furthermore, only 44 of 132 subfields could be analysed, as the others did not meet a minimum threshold of cited papers to facilitate robust analysis. Thus, while this study’s findings are valuable and shed light on a fundamental policy assumption in need of evidential support, they are also extremely preliminary, requiring much more robust assessment before they should be considered sufficient for the basis of future policymaking—robust both in the sense of repeating these analyses on larger samples and in the sense of approaching these phenomena from other angles.

Acknowledgments

This study was performed as part of a larger project on the development and application of a methodological framework guiding the design and implementation of data mining projects in the R&I policy context. The project was commissioned by the European Commission, DG Research and Innovation, Unit A4 – Analysis and monitoring of national research policies.
(Framework Contract: No 2010/S 172-262618). Any use, right of dissemination or copyright belongs to the European Commission. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Commission.

References


Campbell, D., Deschamps, P., Côté, G., Roberge, G., Lefebvre, C., & Archambault, É. (2015). Application of an “interdisciplinarity” metric at the paper level and its use in a comparative analysis of the most publishing ERA and non-ERA universities. Presentation at the 20th International Conference on Science and Technology Indicators (STI), Lugano, Switzerland.

Campbell, D., Tippett, C., Struck, B., Lefebvre, C., Côté, G., St-Louis Lalonde, B., Ventimiglia, A., Roberge, G., & Archambault, É. (2016). Data mining on key innovation policy issues for the private sector: application report. Report prepared for the European Commission, 228 pages (see Section 3.1). Because the report has not yet been published, it must be accessed through a request made to DG Research and Innovation, Unit A4.


Measuring and Forecasting Technology Market Potential

Luan Chunjuan
julielecj@163.com
Dalian University of Technology, Dalian (China)

Abstract
This paper aims at exploring the methodologies and indicators on measuring and forecasting technology market potential, which will benefit us in understanding technology developmental and competitive frontiers and promote patent industrialization and commercialization. Four-quadrant analysis based on patent portfolio data is employed, and two aggregative indicators of disruptive potential and technology maturity are selected, to conduct the study. Empirical study by choosing an emerging technology called cell Immunotherapy in medical domain is investigated from four fields with the filter function embedded in Relecura IP Platform, i.e., Field of Technologies, Field of Sub-Technologies, Field of IPC Codes and Field of CPC Codes. The empirical study help us well understand the situation of technology market potential for a specific technology area from various levels.

Conference Topic
Indicators
Patent Analysis

Keywords
Technology market potential; disruptive potential; technology maturity; four-quadrant analysis; patent portfolio; measuring and forecasting

1 Introduction
The research objective of this study is to explore the methodologies and indicators on measuring and forecasting technology market potential, via four-quadrant analysis based on patent portfolio data, and two aggregative indicators of disruptive potential & technology maturity being selected. Investigations on measuring & forecasting technology market potential will benefit us in deploying emerging technology strategies and selecting R&D investment areas, and it can also afford significant decision-making references for S&T managers and intellectual property (IP) market dealers.

Studies pertinent to the topic of technology market potential have been paid attention to from 1980s by scholars. An approach for estimating the market potential of new energy-saving technologies is presented (Roberts and Greene 1983). This approach is based on the assumption that consumers implicitly minimize total expected lifetime costs when purchasing new technologies and an important result of this approach is that technologies can have zero market potential and that these technologies can be identified based solely on technology-specific information, not on attributes of the consumer population.

It is recognized that understanding technology market potential is not only the premise and foundation of technology trading activities, such as transformation and potential licensing (Bagchi and Mukherjee 2014, Mukherjee 2014), but also is significant basis for R&D researchers to select whether entering a specific technology domain (Chun and Mun 2014, Duchene, Sen et al. 2015). Technology market potential is a crucial indicator for industrial competitors’ capital operating (Svensson 2013), technology trading (Galasso, Schankerman et al. 2013), and market developing trend forecasting (Bayar, Chemmanur et al. 2011). At the same time, technology market potential analysis is also the foundation for companies to implement IP strategies (Engel 2011), to strengthen patent portfolios (Cockburn and MacGarvie 2011, Ismail, Omar et al. 2011), to develop new products (Schmiele 2012), and to exploit emerging technologies (Jun, Yeom et al. 2014).
A patent portfolio is an aggregation or collection of patents owned by a single owner, particularly a corporation, and these patents might have been developed individually by the owner or might have been purchased from the original patent owner. A patent portfolio can be a valuable asset to its owner (Ernst 1998, von Graevenitz, Wagner et al. 2011). The value of a corporation's patent portfolio can be a significant fraction of the overall value of the corporation (Narin 1995, Grimaldi, Cricelli et al. 2015, Song, Seol et al. 2016). Patent portfolio analysis is a methodology employed to evaluate and forecast the developing trend of an innovator or a specific technology field, by constructing the aggregation indicators which can be used to measure patent potential values, and this methodology can also be applied in IP strategies formulating and implementing (Blind, Cremers et al. 2009, Grimaldi, Cricelli et al. 2015, Tripathi, Jana et al. 2016).

Narin proposed a patent portfolio indicator employed to evaluate companies’ R&D performance based on cross-licensing in 1995 (Narin 1995); Ernst put forward 4 models of patent portfolios, i.e., a patent portfolio on the level of technological field, on the company level, on the inventor level, and on the integration of patent-market portfolios (Ernst 1995, Ernst 1997, Ernst 1998). In addition, such topics as the necessity of patent portfolios (Han 2015), the security of patent portfolios (Risch 2013), the function of patent portfolios (Lichtenthaler 2012) and value evaluation of patent portfolios (Schubert 2011), et al., have also been explored.

Technology disruptive potential represents the revolutionary transformation of emerging technologies comparing to the state of the art, and it can change the current technology developing patterns due to its higher innovative index and impacting upon the subsequently technologies significantly (Dereli and Altun 2013, Relecura 2016). Such subjects as the definition of disruptive potential (Ganguly, Nilchiani et al. 2010), future-oriented technology disruptive potentials (Cagnin, Havas et al. 2013), assessment procedures on disruptive potentials of candidate technologies (Dereli and Altun 2013), et al., have been investigated.

Technology maturity can be measured from such dimensions as developing period of a specific technology field, the amount of patents, global progress of the technology field and the extent of its applications (Roussel 1984, Relecura 2016), et al.. Related topics have been investigated, such as the evaluation modes of technology maturity (Rocha 2012, Waring 2015), assessment methods (Nogales-Bueno, Hernandez-Hierro et al. 2014, Nogales-Bueno, Ayala et al. 2015), various risks in the maturation process (Mishra, Das et al. 2016), and the technology maturity of a specific domain (Roussel 1984), et al..

The extant studies have formed the essential foundation of this paper, and have significant referring values for us. Existing indicators in the current four-quadrant analysis tend to be unicity, such as patent activity is represented only by R&D expenditure in a specific technology field. If more factors, such as patent applications, patent citations, and even innovative index, et al. could be taken account into a measuring indicator, the aggregative indicator will be much more convincing, and the analyzing results concluded will be much closer to the real technology competitive market.

Two aggregative indicators including technology disruptive potential and technology maturity will be employed in measuring technology market potential in this study. Relevant influence factors have been taken account into the two indicators. The emerging technology domain of cell immunotherapy is selected as an empirical study, and Relecura IP Platform is selected to conduct the study, exploring the methods and indicators of technology market potential.
measurement, expecting this study support decision-making in the area of emerging technology developing strategy and the S&T achievements transformation.

2 Methodologies & indicators, data source

2.1 Methodologies & indicators

Four-quadrant analysis based on patent portfolio data which is designed by Relecura innovation team (Relecura 2016), is employed in this study. Relecura is a new generation patent search tool. It uses a knowledge discovery framework to simplify patent searches and patent portfolio analysis. Its intuitive interface and unique features provide valuable insights into patents and patent portfolios. This methodology provides an indication of the relative position of the competitors in graphical form. The position of the bubbles in the graph helps us determine how much potential a technology has in the market. The size of the bubble represents the number of all patents worldwide for a specific technology. The Relecura IP Professional Platform analyses the position of the competitors with respect to Technologies/Sub-technologies/Class codes, and positions the values in a four-quadrant format. Four-quadrant analysis is typically applied to display a visual comparison of various parameters, positioned along appropriate X-Y axes, to give an intuitive comparison snapshot of the chosen parameters.

Two indicators, disruptive potential along with X-axis and technology maturity along with Y-axis (Fig. 1), developed by Relecura innovation team (Relecura 2016), are selected to measure technology market potential for a specific technology area.

![Figure 1. Analyzing Model of technology market potential](image_url)

The X-axis, Disruptive potential, indicates the potential of a technology with respect to the IP portfolio at hand (Relecura 2016). It takes into account patent filings, citation counts and innovation index that are indicative of the novelty of a technology. A positive disruptive potential indicates the capacity of the technology to impact the current IP portfolio.

The Y-axis, Tech maturity, indicates the extent to which the technology is prevalent in the IP market (Relecura 2016). It measures the development stage of the technology with respect to the current IP portfolio and includes the number of patents in the technology area, the amount of time the technology has been addressed by the IP portfolio, the global reach of the technology and the implementation or usage measure for the technology.

When a specific technology positions in the first quadrant like T1, it indicates that this technology has a positive disruptive potential and its technology maturity reaches a certain degree. In such a situation, it can be concluded that this technology has a strong market potential with an advantage value in IP business market. If a specific technology locates in the
second quadrant like T2, it indicates that this technology has no disruptive potential, though its technology maturity has reached some degree. Such technologies have little market potential and business value. T3 locates in the third quadrant, shows that such a technology has neither disruptive potential nor technology maturity, and it probably won’t have any market potential. Technologies as T3 have little market potential and business value. Technologies position in the fourth quadrant as T4, reveal that they have technology disruptive potential, and simultaneously, they are still in initial, even infancy developing phase. Such technologies have some uncertainty in their developing process: one possibility is that they can develop gradually and will have great market potential and business value; another possibility is that they cannot develop gradually, or even in the infancy phase for a comparative long period.

2.2 Data source

Cellular immunotherapy is an emerging technology in the medical area. It offers novel, safe and effective routes to treating cancer (Qian, Wang et al. 2016). It is recognized a promising therapy in treating cancer in the recent years (Hochberg, El-Mallawany et al. 2014). The developing trends both of patent filings and granted patents worldwide are shown in Figure 2.

![Figure 2. Developing trends of patent filings and granted patents of Cellular Immunotherapy.](image)

The earliest two patent filings in the technology area of Cellular immunotherapy can be dated back to the year of 1988 with priority in 1987. Only a few patent filings at the beginning phase, 10 patent filings totally before the year of 1993; no patent grants are found before 1991; the first granted patent appears in 1992. The number of total patent filings is 3104 and the total granted patents are 1168, up to our searching date of April 20, 2016. There are 853 patent assignees total, and the top five ones are University of Texas with 114 patent filings, US National Institutes of Health (NIH) with 101 patent filings, Synta Pharmaceuticals with 100 patent filings, US Health and Human Services, abbr. as HHS hereafter, with 90 patent filings, Celera Corp. with 68 patent filings, respectively.

Empirical study: measuring & forecasting technology market potential

Four-quadrant analysis can be employed to conduct the analytic work from the perspectives of assignees, technologies, sub-technologies, and technology class codes including IPC Codes / CPC Codes and US Classes. Here, the filters embedded in Relecura Platform of technologies, sub-technologies, IPC Codes and CPC Codes are selected to do the empirical study.

3.1 Field: Technologies

Firstly, the Field of Technologies has been selected to measure and forecast technology market potential in the area of Cellular immunotherapy by employing four-quadrant analysis embedded in Relecura IP Platform, and we got the map of Fig. 3.
Fig. 3 reveals that in the Field of Technologies, as far as the indicator of Disruptive potential is concerned, the technology of Medicinal Preparations with 2712 patent filings has the strongest disruptive potential, indicating this technology is of absolute synthetic competition advantages in terms of patent filing activities, patent citations and technology novelties. Another three technologies have some degree disruptive potential when it comes to their patent portfolios: Micro organisms/Enzymes with 1,893 patent filings; Therapeutic activity of chemicals and medicinal preparations with 1,469 patent filings; Peptides with 1,461 patent filings, respectively. Technologies locating in the third quadrant, such as Heterocyclic Compounds, et al., nearly have little technology disruptive potential.

![Figure 3. Four-quadrant analysis of technology market potential in the Field of Technologies.](image)

When it comes to the indicator of Tech maturity, it is pretty high of the technology of Medicinal Preparations, indicating that this technology is comparatively prevalent in the current IP market, and it has progressed into a higher developing stage in terms of its amount of patents, the amount of time the technology has developed in the IP portfolios, the global reach of the technology and the implementation or application measure for the technology.

Synthesizing the two indicators both disruptive potential and technology maturity, the technology of Medicinal Preparations has not only stronger technology disruptive potential, but a pretty higher technology maturity, so it can be concluded that this technology has the most promising technology market potential. Though another three technologies, such as Micro organisms/Enzymes, Therapeutic activity of chemicals and medicinal preparations, Peptides, have some extent disruptive potential, they are still situated in a comparatively early developing phase, far away from technology maturity, so they will probably have a long progressing space. Technologies such as Heterocyclic Compounds, et al., having neither disruptive potential nor technology maturity, are of little technology market potential at present.

3.2 Field: Sub-Technologies

Secondly, the Field of Sub-Technologies has been selected to measure and forecast technology market potential in the area of Cellular immunotherapy. It can help us well understand the disruptive potential and technology maturity of various sub-technologies from a perspective of comparative micro level, and further insight their technology market potential. Selecting the field of sub-technologies of Cellular immunotherapy, also by applying four-quadrant analysis embedded in Relecura IP Platform, and we got the diagram of Fig. 4.

The sub-technologies appeared in Fig. 4 have various degree of disruptive potential, with different developing phases in terms of technology maturity. The sub-technology of
Medicinal preparations containing antigens or antibodies with 1,809 patent filings appears to be of comparative advantages for the two indicators, both disruptive potential and technology maturity, so it can be considered the most promising market potential among all. Other sub-technologies, such as undifferentiated human/animal or plant cells with 1,260 patent filings, Antineoplastic agents with 1,224 patent filings, have some degree disruptive potential whereas still being in the initial developing phase, and they will probably have a promising developmental space and even a strong market potential. Still some sub-technologies, such as Medicinal preparations with genetic material-gene therapy with 517 patent filings, et al., they are of some extent disruptive potential yet being with a comparative undeveloped phase in terms of technology maturity, their future are of a certain degree of uncertainty.

![Four-quadrant analysis of technology market potential in the Field of Sub-Technologies.](image)

3.3 Field: Technological Class Codes

1) Field: IPC Codes

The International Patent Classification (IPC), established by the Strasbourg Agreement, provides for a hierarchical system of language independent symbols for the classification of patents and utility models according to the different areas of technology to which they pertain. The appropriate IPC symbols are indicated on each patent document, of which more than 1,000,000 were issued each year in the last 10 years. Choosing the field of IPC Codes of Cellular immunotherapy, also by employing four-quadrant analysis embedded in Relecura IP Platform, and we got the diagram of Fig. 5.

Fig. 5 reveals that the IPC Code of A61P 35/00, representing Antineoplastic agents with 1,201 patent filings, has both a stronger disruptive potential and a higher technology maturity, so it can be concluded that this IPC Code is of stronger market potential among all appearing in Fig. 5. Another two IPC Codes, A61K 39/00 representing Medicinal preparations containing antigens or antibodies with 779 patent filings, A61K 39/395 representing Antibodies; Immunoglobulins; Immune serum with 618 patent filings, respectively, have some degree disruptive potential, yet they just step into the infancy developing stage, still need a comparative long time to technology mature phase, and it is supposed that such IPC Codes could probably get a stronger market potential after a period of development. Still some IPC Codes locating in the fourth quadrant, such as A61K48/00 representing Medicinal preparations containing genetic material which is inserted into cells of the living body to treat genetic diseases / Gene therapy with 468 patent filings, they are of some extent technology disruptive potential, yet they are in the very immature developing phase, existing some
uncertainties, and some of them will probably be of great market potential after a comparative long period of development.

Figure 5. Four-quadrant analysis of technology market potential in the Field of IPC Codes.

2) Field: CPC Codes
Cooperative Patent Classification, CPC, is the new classification system introduced since January 2013 in order to standardize the classification systems of all major patent offices (Montecchi, Russo et al. 2013). CPC has been jointly developed by the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO). The CPC is substantially based on the previous European classification system (ECLA), which itself was a more specific and detailed version of the International Patent Classification (IPC) system (Angue, Ayerbe et al. 2014, Leydesdorff, Alkemade et al. 2015, da Silva, Pinheiro et al. 2016). Here, we choosing the field of CPC Codes of Cellular immunotherapy, also by employing four-quadrant analysis embedded in Relecura IP Platform, and we got the diagram of Fig. 6.

Figure 6. Four-quadrant analysis of technology market potential in the Field of CPC Codes.

Fig. 6 shows that the CPC Code of A61K 39/0011, representing Cancer antigens with 594 patent filings, has a stronger disruptive potential with some degree technology maturity, so this CPC Code is of a comparative stronger market potential among all appearing in Fig. 6. A piles of other CPC Codes, such as CPC=A61K 2039/5158 (Antigen-pulsed cells e.g. T-cells with 408 patent filings) et al., are gathering around X axis, revealing that these groups of technology domains denoted by the CPC Codes have a certain of disruptive potential, yet most of them are still in a situation of initial developing phase, even just in a infancy stage. Some of them will probably develop into a mature stage gradually and will be of disruptive
potential, and some of them may be not. Another CPC Code, A61K 35/17, denoting Lymphocytes; B-cells; T-cells; Natural killer cells; Interferon-activated or cytokine-activated lymphocytes with 263 patent filings, positioning in the fourth quadrant, nearing original point, neither has it disruptive potential nor technology maturity at present, and it needs time to predict its market potential.

4 Conclusions and discussions

4.1 Conclusions

This study has shed light on how to measure and forecast technology market potential. It will promote technology commercialized if the information of technology market potential could be measured, forecasted and understood. We proposed the two aggregative indicators, disruptive potential and technology maturity, embedded in the IP Professional Platform Relecura, an international intellectual property Platform, to measure and forecast technology market potential. These two indicators are designed based on patent portfolio data, and each of them is a combination of multiple factors taken into account. The indicator of disruptive potential has been designed by considering such factors as innovators’ patent activities, patent citations and innovative index representing technology novelty, et al.; the other indicator of technology maturity has been designed according to the factors, such as the amount of patents in a specific technology domain, the duration of a technology has been developing, and the level of a technology reach and its implementation, et al..

We also conduct an empirical study on measuring and forecasting technology market potential of an emerging technology named cell Immunotherapy in medical domain, by employing the four-quadrant analysis based on patent portfolio data, and selecting the two aggregative indicators, disruptive potential and technology maturity, from four fields with the filter function embedded in Relecura IP Platform, that is, Field of Technologies, Field of Sub-Technologies, Field of IPC Codes and Field of CPC Codes. The empirical study help us well understand the situation of technology market potential for a specific technology from various levels.

Finally, further research is necessary to investigate technology market potential from a comparatively micro level, for example, individual patent level, instead of a technology area. It will help patent assignors or assignees much if they can get the information of technology market potential of their patents pending to be commercialized or industrialized.

4.2 Discussions

The innovation of this study lies in, the two aggregative indicators, disruptive potential and technology maturity, taking account into multiple factors, are of much stronger functions in measuring and forecasting technology market potential, comparing to the extant results. It affords us a big inspiration that the analysing four quadrant model based on patent portfolio analysis proposed by previous researchers (Ernst 1998, Santiago, Martinelli et al. 2015). The design of the indicator of disruptive potential, considering not only patent portfolios of assignees, but synthesizing technology impacts on sequent technology developments, and even more, this indicator combines the analysis of technology novelty characteristics by adding innovative index. More disruptive potential, more destructive force, more challenging power for the existing technology market, and such disruptive potential technologies are likely to transform the current technology market developing patterns.
The indicator of technology maturity, taking into account a series of factors including the amount of patents and its developing time, also the level it reaches and the usage globally (Relecura 2016). In such a four-quadrant analysis model, we can predict a specific technology will have a promising market potential in the future, if it is of a strong disruptive potential, and simultaneously it still locating in a very beginning stage, or even in a infancy phase. A technology will probably step into a declining phase next, if it has strong disruptive potential and already it is in a comparative technology developing mature stage, then its advantageous market potential probably won’t last long. If a technology has little disruptive potential, then it is supposed to be of little market potential, either, regardless of its technology maturity.

Indicators employed in this study, not only can be applied by the national planning department in deploying emerging technology developing strategies, but also can be adopted by corporations in selecting R&D projects. At the same time, the indicators could be chosen by business investors for decision making. It will help us understanding the developing trend and direction of the frontier technologies, improve our decision-making quality in selecting new technology and implementing intellectual property strategy, and also promote patents commercialization. It can be easily imagine that if an innovator chooses a technology with great disruptive potential, and it is still in infancy stage, then such a technology is probably developed successfully in the near future, and it will have great market competitive advantages, further it will be easily commercialized for its great market potential. Due the limit of pages, the rationale on constructing the indicators will be illustrated in the future studies.

Acknowledgments
I do appreciate Mr. George Koomullil, the CEO of Relecura IP Professional Platform Inc. and the Director Sheena Mathew, for their facilitating me much in my employing the program of Relecura. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 71473028; Project of the National Social Science Fund under Grant No. 14BTQ030. Fundamental research funds for the central universities under Grant No. DUT17RW224.

References


Detecting Science & Technology of Lockheed Martin Corporation throughout Funded Papers

Zhao Rongying\(^1\) Liang Zhisen\(^2\)

\(^1\)zhaorongying@126.com
Research Center for Chinese Science Evaluation, Wuhan University, Wuhan (China)
School of Information Management, Wuhan University, Wuhan (China)

\(^2\)625928798@qq.com
Research Center for Chinese Science Evaluation, Wuhan University, Wuhan (China)
School of Information Management, Wuhan University, Wuhan (China)

Abstract
As one of research funding providers, private company will influence development of science and technology throughout its support. While funded papers are mainly used to test their impacts of funding in recent researches, it can be expanded to detect the intellectual structure and scientific frontier of this provider. This study examines the co-word network of scientific papers funded by Lockheed Martin Corporation. We explore the dynamics of this network from 2009 to 2016 based on publication records collected from Web of Science. In our analysis, we firstly divide records into three main clusters according to disciplines, using HCA and SPSS. Secondly, we use the mixed strategy to select each cluster’s keywords, which are extracted by BibExcel before. Lastly, we map the network to detect keywords’ communities and core sub-network in each cluster, using Gephi and K-cores decomposition method. The analytical results demonstrate that what this corporation funded is related to their business and research areas, and detailed scientific fields are founded.

Conference Topic
Social network analysis, Co-occurrence analysis

Keyword
Funded paper; Co-word analysis; Hierarchical cluster analysis; Social network analysis; BibExcel; SPSS; Gephi; intellectual structure; scientific frontier

Introduction
As a key resource in current academic community, research funding has an irreplaceable function in research development (Vardakas et al., 2015; Ebadi and Schiffauerova, 2016; Zhao et al., 2016), as well as fulfills researchers’ hope to expedite and disseminate high-quality research (Zhao, 2010). Common sources of research funding include the researchers themselves, their home institutions in the form of salary and research infrastructure, and grants internal to the researchers’ institutions or, more importantly, by public or private funding agencies outside the researchers’ institutions (Zhao, 2010). Besides financial support, research funding sometimes would be seen as physical support, such as device and facility, especially from company or laboratory.
On the other hand, research funding is not an act of kindness (Li, 2011). Researchers are confronted with an environment of decreasing funding resources (Hoyla et al., 2016), because funders make detailed demands on the target, the purpose of the project and the attribution of research achievements, as well as lead to discussion of how to make financial resources go further, and to the concern that some researchers take more money from them for a particular project than needed (Rigby and Julian, 2014). Taking the National Natural Science Foundation of China (NSFC) as example, it had an acceptance rate of 22.87% in General Program and 22.00% in Key Program in 2016 (National Natural Science Foundation of China, 2016). In other words, more than 75% of the proposals were rejected. They just did not conform to what NSFC preferred, or rigorous competition-based funding policy, consisting of procedures, practices, stipulating grant conditions and regulatory criteria (Braun, 1998; Wang and Shapira, 2011). To put it another way, with limitations, a funder can indirectly affect development of science and technology (S&T).

Since funder can influence the development of scientific fields, funded papers, one of the most important outcome of researches, may contain information of what funder see development of S&T as. However, these papers are usually used to evaluate the efficiency of financial support in scientometrics, ignoring this information. Considering co-word analysis may be a good way to map the intellectual structure of academic researches, as well as find frontiers of S&T, we take co-word analysis to detect the information of funder.

Public funding agency, private foundation and company are common research funding providers. In this paper, we chose Lockheed Martin Corporation (Lockheed) as our example. As an US corporation with worldwide interests, this corporation organizes or participates in numbers of individual research programs and joint projects, providing researchers financial support. In addition, Sandia National Laboratories (Sandia), one of the national laboratories of US Department of Energy, operated by Sandia Corporation, a Lockheed Martin company, also provides financial or physical support. As a part of Lockheed from 1993-2017, it is no wise to exclude Sandia and its funded researches. Therefore, we treat Lockheed and Sandia as an integrated whole (LMT), and funded paper includes which paper receives financial or physical support.

This paper investigated the co-occurrence network of keywords form scientific journal articles funded by LMT, attempting to map the intellectual structure of researches and find frontiers of S&T supported by this corporation. We explore the funded articles of this network from 2009 to 2016 based on large data derived from Web of Science, Science Citation Index Expanded (WoS). It is organized as follows: after giving a brief review of assessment in funded researches, we describe in section “Data and Methodology”. Then, we present and analyze the detailed results in ‘Results and discussion’ section, followed by a conclusion section including the major findings and limitations.

Literature

With the help of quantitative tools provided by bibliometric and scientometrics, analysis of funded researches has become a hotspot in recent years.

Firstly, Acknowledgement is the main source to identify which research is funded, and collect funding information, such as funding program, agency, and grant number. Wang and Shapira
(2011) used a novel bibliometric approach in the context of nanotechnology research, to undertaking funding acknowledgement analysis which links research outputs with their funding sources.
Secondly, it is usually to find out difference between funded and non-funded researches in the same field, as well as funding effect in different countries. Zhao (2010) compared grant-based funding of research with non-grant-based research, both published in library and information science journals. And compared Canada with the United States in terms of the impact of public grants and scientific collaborations on subsequent nanotechnology-related publications, Tahmooresnejad et al. (2015) found that while research funding had a significantly positive linear impact in Canada and a positive non-linear impact in the United States on the number of papers, and it just only yielded a positive impact in the US in terms of citations.
Also, funded paper can be used to evaluate an academic journal, such as Sun (2010) evaluated Information Science, by the ratio of funded paper, the scale of the grant, subjects, collaboration, and institution the authors belong to, and Wei (2011) investigated funded papers in Library Tribune published from 2006 to 2009 by quantitative methods.
Thirdly, choosing funding agency and program is another point. For example, Gaughan and Bozeman (2002) examined effects of funding from the National Science Foundation (NSF), Gaughan (2009) chose the National Institute of Health (NIH) to compare impact on the paper publication rate adjusted by age, Ida and Fukuzawa (2013) investigated the 21st Century Centers of Excellence (COE) Program in Japan, and then Rigby and Julian (2014) indicated that papers achieved a higher citation count, funded by European Molecular Biology Organization (EMBO) and the Human Frontier Science Program (HFSP) at the same time.
Based on assessment results of funded researches, mainly using the number of papers and citation per paper as indicators, agency can watch on the situation and patterns of these investments and related outputs (Zhao et al., 2016), as well as decide on subsidizing potential domains, institutions, or individual, which thus enhances the efficiency of research resource allocation to achieve maximum production results (Liaw et al., 2014). What’s more, canonical funding agency or program, established by government, is the most popular choice for academia.
In addition, besides methods of scientometrics, there are some other ways to analyze funded research. Hoyla et al. (2016) used Proposal-Evaluation-Grant System (PEGS), an agent simulation system that models individual researchers in their activities related to getting funding, aiming for promotions and making science, to simulate different research funding allocation policies. Comins (2015) applied a data-mining approach to dig up technological importance of government-funded patents in the private sector.
Last but not least, co-word analysis were used to map the intellectual structure, or identify topics of the research denoted by funding agency or program, such as NOAA’s Office of Ocean Exploration and Research (OER) (Belter, 2013), and “The impact of economic transition on Central and Eastern European enterprises” (Topalli et al., 2016). However, they just studied on public program, not a company, especially scientific and technological enterprise which denote much to and participate in government researches or university studies. Therefore, with co-word analysis, this study can be extended to private company.
Data and Methodology

Data Collection

WoS was used as source of data, which catches attention of scientific and technological academia in the world and is the approach to top level and high quality of basic science fields (Kang and Su, 2008). It has started implementation of Funding Acknowledgements since August 2008. The funding acknowledgement fields contain data on research funding agencies, grant award numbers and associated funding acknowledgement text reported by paper authors (Thomson Reuters 2010; Wang and Shapira, 2011). In order to collect scientific journal articles produced with LMT funding from 2009 to 2016, we searched the “FO = Funding Agency” field tag.

The first search string was “Lockheed Martin”, receiving 855 records. Analyzing their detailed funding agencies, we found that there were some articles also funded by Sandia, which belongs to LMT until May 1, 2017.

The second search string “Sandia Co* OR Sandia National Lab*” was executed and we collected 2,012 records.

Considering some affiliations of Lockheed Corporation or Martin Marietta, the predecessors of Lockheed, may still exist, we respectively searched both of them in next. Compared with the first search string, string “Lockheed” received another ten records. And analyzing the string “Martin Marietta”, only “Martin Marietta Energy Systems” belongs to Lockheed Martin.

The ultimate search string used in this method was followed:

“FO = ((Lockheed OR Martin Marietta Energy Systems) OR (Sandia Corp* OR Sandia Com* OR San Cor* OR Sandia National Lab*))”

This search was executed and obtained 2,758 records. We downloaded and examined them in full-text format. In these records, some are lack of “DE (Author Keywords)” or “ID (Keywords plus)” fields. The situation of articles per year is shown in Table 1. Reducing those lack of both DE and ID, 2,690 articles were finally selected.

<table>
<thead>
<tr>
<th>Publication year</th>
<th>Articles</th>
<th>Articles lack of DE</th>
<th>Articles lack of ID</th>
<th>Articles lack of both</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>289</td>
<td>122</td>
<td>19</td>
<td>8</td>
</tr>
<tr>
<td>2010</td>
<td>316</td>
<td>160</td>
<td>30</td>
<td>14</td>
</tr>
<tr>
<td>2011</td>
<td>386</td>
<td>189</td>
<td>25</td>
<td>9</td>
</tr>
<tr>
<td>2012</td>
<td>394</td>
<td>181</td>
<td>20</td>
<td>8</td>
</tr>
<tr>
<td>2013</td>
<td>365</td>
<td>177</td>
<td>27</td>
<td>10</td>
</tr>
<tr>
<td>2014</td>
<td>376</td>
<td>166</td>
<td>26</td>
<td>8</td>
</tr>
<tr>
<td>2015</td>
<td>343</td>
<td>163</td>
<td>27</td>
<td>9</td>
</tr>
<tr>
<td>2016</td>
<td>289</td>
<td>112</td>
<td>14</td>
<td>2</td>
</tr>
</tbody>
</table>

Methods and Tools

In this paper, we use the co-word analysis, including cluster analysis and social network
analysis. Co-word analysis is a technique that uses pattern of co-occurrence of words and phrases in a corpus (Ravikumar, 2015). Combining the bibliometrics and text mining techniques, it can reflect the relation between terms in order to mine the deeper meaning of the literatures (An and Wu, 2011). Especially, occurring of two keywords within the same paper indicates a relationship between the topics to which they refer (Cambrosio et al., 1993). With the co-word analysis, it is possible to map the intellectual structure of academic researches, as well as find frontiers of S&T, focused on by LMT.

Cluster analysis refers to a class of data reduction methods used for sorting cases, observations, or variables of a given dataset into homogeneous groups that differ from each other (Yim and Ramdeen, 2015), and facilitating comparison among similar entities and support decision making (Tahmooresnejad et al., 2015). From cluster analysis, as a powerful tool for comparison of entities and for high-lighting inequalities (Tashobya et al., 2016), hierarchical cluster analysis (HCA) is used to divide records into several subgroups based on disciplines, carried out by the SPSS statistical software (version 22).

Compared with other approaches, social network analysis are better for conducting co-word analysis (Ding et al., 2001; Hu et al., 2013). From social network analysis, we use both BibExcel and Gephi (version 0.9.1) with the former providing a file with co-word matrix for the latter. As an application software that can analyze information from WoS publication records (Jiang and Chen, 2010), BibExcel extracts keywords form the publication records, counts their times of occurrence, puts in numerical order, and finally generates Pajek NET files with co-occurrence matrices of keywords. Gephi is an open-source dedicated network analysis software used in data visualization (Calma and Davies, 2016), supporting Pajek NET files in after mapping network of keywords.

Step wise procedures of analyses are mentioned below as it is shown in Figure 1:
Preliminary measurement and normalization

The selected records contained two kinds of keyword fields, “DE” and “ID”. We merged “DE” with “ID” into a new field, named “R-DE”. Extracting keywords in “R-DE” with BibExcel into Excel, we found three problems:

P1. In one article, there are several terms for the same thing. “DE” is a field that author pluses into the article, while “ID” is produced by WoS, based on the article. The combination makes it likely that the same thing may have more than one term in one article. For example, “microelectromechanical systems” and its abbreviation “mens” occurred simultaneously at the times in one record.
P2. The same thing has different spellings among different articles. It includes singular and plural, American English and British English, phrases with and without hyphen, as well as popular name and its scientific name or chemical formula.

P3. The keywords are too scattered as a whole. Different form articles with the same subject, articles selected in this paper just have the same funding agency, LMT, and belong to several different disciplines. Therefore, keywords in this paper gather on several subjects.

In order to resolve P1 and P2, we firstly created a keywords table (see Table 2) to normalize the extracted keywords with Excel. After unifying terms and spellings of the identical thing, deleting duplicated terms in the same article, and then modifying “R-DE” fields in the original records, we finally got new modified records with 11,169 keywords, less than 18,386 originally had.

<table>
<thead>
<tr>
<th>Raw Keywords</th>
<th>Normal Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>chemical vapor deposition</td>
<td>chemical-vapor-deposition</td>
</tr>
<tr>
<td>chemical-vapor-deposition</td>
<td>chemical-vapor-deposition</td>
</tr>
<tr>
<td>cvd</td>
<td>chemical-vapor-deposition</td>
</tr>
</tbody>
</table>

Cluster analysis of disciplines

The P3 forced us to divide the whole modified records into several subgroups. There were two fields related to disciplines, “SC” and “WC” in the records. In SCI-Expend, “WC” referred to Web of Science Category, including 176 categories of subjects (Thomson Reuters., 2017). By contrast, “SC” was like first level disciplines, while “WC” was secondary level disciplines. We extracted disciplines of “SC” and “WC” separately with BibExcel, building Table 3. It was shown that there were 137 disciplines in “WC”, divided into 80 disciplines in “SC”. The latter was more general and easier to process than the former. Therefore, with 2,489 publication records, accounting for close on 93% of the whole modified records, the top 15 “SC” disciplines were took to build a co-occurrence matrix (see Table 4).

However, the co-occurrence frequency was absolute value, so it was difficult to reflect real interdependence between words, in the use of statistical software for cluster analysis (Lu, 2016). For this, the matrix was converted to correlation matrix with Ochiai coefficient (see Table 5), and then dissimilarity matrix, using the method of subtracting elements of 1 converted it to dissimilarity matrix (see Table 6) (Li and Wu, 2011).

Ochiai coefficient could numerate as the following:

\[ Ochiai = \frac{c_{ij}}{\sqrt{c_{i} \times \sqrt{c_{j}}}} \quad (1) \]

\( c_{ij} \) is regarded as time of co-occurrence between keyword i and keyword j, \( c_{i} \) is regarded as time of keyword i and \( c_{j} \) is time of keyword j.

In dissimilarity matrix, bigger the element was, farther the distance and the lower correlation between the keywords were (Zhang and Ma, 2007). The matrix was imported into SPSS to conduct HCA, using Ward’s method and squared Euclidean distance (Dehdariad et al., 2014),
with no standardization. Finally, HCA was illustrated using a dendrogram (Figure 2).

Table 3. The Contrast between “SC” and “WC” of the whole modified 2,690 records (section)

<table>
<thead>
<tr>
<th>“SC” Field</th>
<th>“WC” Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physics</td>
<td>Physics, Applied</td>
</tr>
<tr>
<td></td>
<td>Physics, Condensed Matter</td>
</tr>
<tr>
<td></td>
<td>Physics, Atomic, Molecular &amp; Chemical</td>
</tr>
<tr>
<td></td>
<td>Physics, Fluids &amp; Plasmas</td>
</tr>
<tr>
<td></td>
<td>Physics, Multidisciplinary</td>
</tr>
<tr>
<td></td>
<td>Physics, Mathematical</td>
</tr>
<tr>
<td></td>
<td>Physics, Particles &amp; Fields</td>
</tr>
<tr>
<td></td>
<td>Physics, Nuclear</td>
</tr>
<tr>
<td>Materials Science</td>
<td>Materials Science, Multidisciplinary</td>
</tr>
<tr>
<td></td>
<td>Materials Science, Coatings &amp; Films</td>
</tr>
<tr>
<td></td>
<td>Materials Science, Characterization &amp; Testing</td>
</tr>
<tr>
<td></td>
<td>Materials Science, Composites</td>
</tr>
<tr>
<td></td>
<td>Materials Science, Ceramics</td>
</tr>
<tr>
<td></td>
<td>Materials Science, Biomaterials</td>
</tr>
<tr>
<td>Chemistry</td>
<td>Chemistry, Physical</td>
</tr>
<tr>
<td></td>
<td>Chemistry, Multidisciplinary</td>
</tr>
<tr>
<td></td>
<td>Chemistry, Inorganic &amp; Nuclear</td>
</tr>
<tr>
<td></td>
<td>Chemistry, Analytical</td>
</tr>
<tr>
<td></td>
<td>Chemistry, Applied</td>
</tr>
<tr>
<td></td>
<td>Chemistry, Organic</td>
</tr>
<tr>
<td></td>
<td>Chemistry, Medicinal</td>
</tr>
</tbody>
</table>

Table 4. The matrix of top 15 “SC” categories (Section)

<table>
<thead>
<tr>
<th></th>
<th>Physics</th>
<th>Materials Science</th>
<th>Chemistry</th>
<th>Engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physics</td>
<td>924</td>
<td>247</td>
<td>222</td>
<td>84</td>
</tr>
<tr>
<td>Materials Science</td>
<td>247</td>
<td>657</td>
<td>303</td>
<td>51</td>
</tr>
<tr>
<td>Chemistry</td>
<td>222</td>
<td>303</td>
<td>569</td>
<td>19</td>
</tr>
<tr>
<td>Engineering</td>
<td>84</td>
<td>51</td>
<td>19</td>
<td>567</td>
</tr>
</tbody>
</table>

Table 5. The correlation matrix of top 15 “SC” categories (Section)

<table>
<thead>
<tr>
<th></th>
<th>Physics</th>
<th>Materials Science</th>
<th>Chemistry</th>
<th>Engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physics</td>
<td>1</td>
<td>0.317013884</td>
<td>0.306168746</td>
<td>0.116051771</td>
</tr>
<tr>
<td>Materials Science</td>
<td>0.317013884</td>
<td>1</td>
<td>0.495568634</td>
<td>0.083559525</td>
</tr>
<tr>
<td>Chemistry</td>
<td>0.306168746</td>
<td>0.495568634</td>
<td>1</td>
<td>0.033450756</td>
</tr>
<tr>
<td>Engineering</td>
<td>0.116051771</td>
<td>0.083559525</td>
<td>0.033450756</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 6. The dissimilarity matrix of top 15 “SC” categories (Section)

<table>
<thead>
<tr>
<th></th>
<th>Physics</th>
<th>Materials Science</th>
<th>Chemistry</th>
<th>Engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physics</td>
<td>0</td>
<td>0.682986116</td>
<td>0.693831254</td>
<td>0.883948229</td>
</tr>
<tr>
<td>Materials Science</td>
<td>0.682986116</td>
<td>0</td>
<td>0.504431366</td>
<td>0.916440475</td>
</tr>
<tr>
<td>Chemistry</td>
<td>0.693831254</td>
<td>0.504431366</td>
<td>0</td>
<td>0.966549244</td>
</tr>
<tr>
<td>Engineering</td>
<td>0.883948229</td>
<td>0.916440475</td>
<td>0.966549244</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 2. Dendrogram of top 15 disciplines showing clusters

Social network analysis of keywords

After clustering, we divided 2,489 publication records into three groups. Cluster “MSCP” had 1,626 publication records, extracted 6,426 keywords with BibExcel, while “ET-EMCM” had 903 records with 5,098 keywords and “IEAPO” had 454 records with 2,455 keywords. With so many keywords, it was no wise to put all of them in Gephi, which might make visualization of network deprived of conciseness and readability. Therefore, a strategy of keyword selection was needed. Yang (2014) had introduced a new strategy that the same number of keywords was selected respectively based on two strategies, one was word frequency threshold and the other was co-occurrence intensity threshold, and then the two selected keyword collections were merged into the final selected keyword collection, which proved to have a better effectiveness of co-word analysis. However, it was difficult to select the same number of keywords in these publication records, so it was compromise to use similar number. Co-occurrence intensity thresholds of three clusters were all set at 3, while word frequency thresholds of “ET-EMCM” and “IEAPO” were 5 but “MSCP” was 6. Finally, three combined keyword collections (see Table 7) came out respectively in Pajek Net file.

Next, keywords of the three clusters were visualized in social networks separately. In Gephi, a
node was a keyword, and an edge showed the relationship between two or more keywords, which referred to co-occurrence of keywords. Importing Pajek Net files, directed graphs set by “Force Atlas” layout algorithms were generated. Counting average path length and modularity, we could get all nodes’ betweenness centrality and closeness centrality, as well as detect communities of keywords. Each cluster has two types of graph. One was communities of keywords, the other was core sub-group of keywords with degree centrality and betweenness centrality. The former interpreted the structure of topics represented by keyword communities, identified by node’s color. The latter was filtered by K-cores decomposition method, which was used for exploring the hierarchy of social network and detecting core sub-groups (Baxter et al., 2012). Setting the highest K-core value, it visualized the lastly filtered nodes and edges. The K-core value of a node, defined as \( k \), described that this node had at least \( k \) interconnections, as well as referred to the core degree of node (Xiao et al., 2016; Zhang et al., 2014). Color of node in latter graph showed the degree, while size of node presented the betweenness centrality in both graphs. The darker color meant the higher degree, and the bigger size meant the higher betweenness centrality.

Table 8 lists the three clusters with number of nodes and edges, modularity, number of communities, average path length, and the highest K-core value in Gephi.

### Table 7. Word frequency thresholds and co-occurrence intensity thresholds, as well as their numbers of combined keyword in “MSCP”, “ET-EMCM” and “IEAPO”

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Word frequency threshold</th>
<th>Co-occurrence intensity threshold</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Threshold</td>
<td>Keywords</td>
<td>Threshold</td>
</tr>
<tr>
<td>MSCP</td>
<td>6</td>
<td>392</td>
<td>3</td>
</tr>
<tr>
<td>ET-EMCM</td>
<td>5</td>
<td>238</td>
<td>3</td>
</tr>
<tr>
<td>IEAPO</td>
<td>5</td>
<td>133</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 8. The number of nodes and edges, modularity, number of communities, average path length, and the highest K-core value of “MSCP”, “ET-EMCM” and “IEAPO” in Gephi

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Nodes</th>
<th>Edges</th>
<th>Modularity</th>
<th>Communities</th>
<th>Avg. Path length</th>
<th>K-core value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSCP</td>
<td>513</td>
<td>10307</td>
<td>0.284</td>
<td>7</td>
<td>2.3127</td>
<td>29</td>
</tr>
<tr>
<td>ET-EMCM</td>
<td>321</td>
<td>3597</td>
<td>0.409</td>
<td>9</td>
<td>2.4211</td>
<td>16</td>
</tr>
<tr>
<td>IEAPO</td>
<td>177</td>
<td>1868</td>
<td>0.347</td>
<td>4</td>
<td>2.2842</td>
<td>18</td>
</tr>
</tbody>
</table>

Results and discussion

Clusters of disciplines

Figure 2 presents an output of the clustering process for the top 15 disciplines as a dendrogram, showing how the disciplines were successively grouped into clusters until they formed a single cluster. A line has been drawn through the dendrogram to illustrate the point at which three clusters were determined. The three clusters are as follows:

“MSCP”: The cluster includes 4 disciplines, such as Physics, Materials Science, Chemistry, and Science & Technology. In addition, except Physics, the rest can be merged into a more
homogeneous sub-cluster, which infiltrate mutually and develop crosswise. MSCP may be the most significant scientific fields LMT payed attention to, which has almost 65% of 2,489 publication records.

“ET-EMCM”: It can be separated into three sub-clusters. The first is composed of Energy & Fuels and Thermodynamics, the next is Engineering and Mechanics, and the last is Computer Science and Mathematics.

“IEAPO”: As the more heterogeneous cluster, it is formed by Optics, Astronomy & Astrophysics, Instruments & Instrumentation, Polymer Science and Electrochemistry. Although the determining point could be located closer to the left side of the plot, illustrating that the clusters being merged became more homogeneous (Yim & Ramdeen, 2015), and they had a number of duplicate recorders with each (see the Venn diagram in Figure 3), it was reasonable for this paper. The number of the clusters was modest for later co-word analysis, and the proportion of duplicated records with another two was less than 35% in each cluster, except IEAPO up to more than 45%. It was still a successful clustering process.

![Figure 3. Venn diagram of relation among three clusters’ records](image)

_Gephi files of MSCP_

There are 7 communities of MSCP in Figure 4, with 0.284 of modularity. Connected with “growth”, most of nodes from Group 0 and Group 1 gather together, and almost all nodes with high betweenness centrality from Group 2, Group 3, Group 4, Group 5 and Group 6 connected with Group 0 and Group 1 (see Figure 5).

1) Group 0 and Group 1 are related to Materials Science. The former is about Coatings & Films, while the latter is Composites. In Group 0, “thin-films”, “graphene”, “chemical-vapor-deposition” are included, which are obviously about thin film, a layer of material ranging from fractions of a nanometer (monolayer) to several micrometers in thickness, playing an important role in the development and study of materials with new and unique properties (Wikipedia, 2017b). Chemical vapor deposition (CVD) is a chemical process used to produce high quality, high-performance, solid materials, which is often used in the semiconductor industry to produce thin films (Wikipedia, 2017a), while Graphene is one kind of thin film. Group 1 has keywords about composite, such as “mechanical-properties”, “carbon nanotubes”, “nanocomposites”, “hybrid composites”.

2) Group 3 and Group 5 are related to the application of molecular dynamics, a computer
simulation method for studying the physical movements of atoms and molecules (Zhou et al., 2016), because “molecular-dynamics” has high betweenness centrality behind “model” in Group 3, and “molecular-dynamics simulations” in Group 5.

3) Group 2 is closely related to study of nanowires, especially its application of battery, with “nanowires”, “ion batteries”, “cathode “, “energy-storage”.

4) Unlike the groups noted above, Group 4 and Group 6 are too decentralized to have unique theme. Group 4 can be divided into two parts, one is centered on “interface”, gathering in the center of Figure 4, while the other takes “plasma” as center, in the edge of Figure 4. Unlike Group 4, Group 6 is also divided into two parts: all nodes with high betweenness centrality flock together, linking to “growth”, and the rest just located on the edge.

Figure 6 shows the core sub-network of $k = 29$ in MSCP, with 102 nodes, and Table 9 lists the top 10 nodes with frequency, betweenness centrality and modularity class. “thin-films” is the node with the highest degree centrality, and “growth” and “nanoparticles” are nodes with the highest betweenness centrality. Obviously, thin-film is the main study object in MSCP, which relates to Materials Science, Physics, and Nanoscience & Nanotechnology. Especially, the growth of thin film, and thin films in nano-scale maybe the major research areas.

Figure 4. The communities of keywords in MSCP
Figure 5. Dynamic network centered on “growth”

Figure 6. Core sub-network of $k = 29$ in MSCP

Table 9. Top 10 core nodes of $k = 29$ in MSCP, with higher Betweeness Centrality, Frequency and Group

<table>
<thead>
<tr>
<th>Node label</th>
<th>Frequency</th>
<th>Betweeness Centrality</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>growth</td>
<td>94</td>
<td>8330.722</td>
<td>0</td>
</tr>
<tr>
<td>model</td>
<td>76</td>
<td>5281.611</td>
<td>3</td>
</tr>
<tr>
<td>Keywords</td>
<td>Count</td>
<td>Frequency</td>
<td>Modularity</td>
</tr>
<tr>
<td>----------------------</td>
<td>-------</td>
<td>-----------</td>
<td>------------</td>
</tr>
<tr>
<td>nanoparticles</td>
<td>59</td>
<td>4667.477</td>
<td>0</td>
</tr>
<tr>
<td>thin-films</td>
<td>183</td>
<td>3237.723</td>
<td>0</td>
</tr>
<tr>
<td>nanostructures</td>
<td>51</td>
<td>3056.285</td>
<td>0</td>
</tr>
<tr>
<td>dynamics</td>
<td>50</td>
<td>2677.626</td>
<td>5</td>
</tr>
<tr>
<td>silicon</td>
<td>59</td>
<td>2391.777</td>
<td>0</td>
</tr>
<tr>
<td>molecular-dynamics</td>
<td>44</td>
<td>2344.556</td>
<td>3</td>
</tr>
<tr>
<td>nanowires</td>
<td>53</td>
<td>2286.063</td>
<td>2</td>
</tr>
<tr>
<td>performance</td>
<td>38</td>
<td>2199.352</td>
<td>2</td>
</tr>
</tbody>
</table>

*Gephi files of ET-EMCM*

Keywords in ET-EMCM are divided into 9 communities (see Figure 7), while the modularity is 0.409. Ignoring the control function of “model” in the network, the communities in Figure 7 are demonstrated as follows, based on their nodes with high betweenness centrality:

1) Group 8 can be divided into three sub-groups. Two groups are mainly related to Energy & Fuels and Thermodynamics, such as study of electrode, and the energy of chemical reaction. For example, the first includes “hydrogen-production”, “electrodes”, “thermal boundary conductance” and so on (see Figure 8), while the second takes “energy” as center connecting “hydrogen”, “heat transfer”, “thermodynamic analysis” (see Figure 9). The last focuses on microelectromechanical systems (MEMS), the technology of microscopic devices widely used in high and new technology industry, merged at the nano-scale into nanoelectromechanical systems (NEMS) and nanotechnology (Wikipedia, 2017c), having “microelectromechanical systems” and “microfluidics” (see Figure 10). Microfluidics is a new development of MEMS, as a multidisciplinary field intersecting engineering, physics, chemistry, biochemistry, nanotechnology, and biotechnology (Wikipedia, 2017d).

2) Group 3, Group 0 and 6 reflect scientific methods. For example, Group 3 is obviously related to the method of fluid dynamics, having “finite element”, “navier-stokes equations”, and “computational fluid dynamics”, especially Navier–Stokes equations describes the motion of viscous fluid substances. And, Group 0 has “model predictive control”, “nonparametric regression”, as well as Group 6 includes “posteriori error analysis”, “sensitivity analysis”, etc.

3) Groups 1, Group 2 and Group 5 are about molecular dynamics, especially molecular dynamics simulation, studying on the growth, strain, crack, plasticity, propagation, elasticity, and degradation of metals or alloys. For example, Group 1 takes “deformation” as center, linking to “molecular-dynamics simulations”, “strain” and “crystal plasticity”, etc. With the node of “molecular-dynamics”, Group 2 also has “propagation”, “elasticity theory”, “crack-growth”, etc. Group 5 has nodes connected to Group 2, such as “shape optimization”, “damage tolerance”, “finite element analysis”, which are connected with Group 2’s “finite-element method” and “elasticity” (see Figure 11). The shape optimization is related to molecular dynamics simulation.

4) Phase, as a physically distinctive form of a substance, such as the solid, liquid, and gaseous states of ordinary matter, is the main theme of Group 4, which is consisted of “isoenergetic-isochoric flash”, “phase-equilibria”, “phase stability”, “liquid equilibrium”, etc.

5) Group 7 also can be divided into two main sub-groups. One is related to Computer Science, Hardware & Architecture, with “memristors”, “resistive random access memory (rram)”, “electrical characteristics”, “resistive memory” (see Figure 12). The other is bound up with the...
application of artificial material, taking “metamaterial” as center linking to “artificial magnetic conductors” and “horn antennas” (see Figure 13). For example, metamaterial is a material engineered to have a property that is not found in nature (Kshetrimayum, 2004), and artificial magnetic conductors is one of them, which is used in antenna.

Figure 14 shows the core sub-network of \( k = 16 \) in ET-EMCM, with 66 nodes, and Table 10 lists the top 10 nodes with frequency, betweenness centrality and modularity class. Keywords with high betweenness centrality over 1,000 are mostly about methods, such as “finite-element method”, “finite element”, “equations”, “simulation”, “optimization”, and, “formulation”. It is obviously that the core network is about scientific methods. One of the most typical methods in this field is molecular dynamics, detecting the property of given materials.

Figure 7. The communities of keywords in ET-EMCM
Figure 8. Dynamic network centered on “generation”

Figure 9. Dynamic network centered on “energy”
Figure 10. Dynamic network centered on “microelectromechanical systems”

Figure 11. Dynamic network centered on “shape optimization”

Figure 12. Dynamic network centered on “impact”
Figure 13. Dynamic network centered on “metamaterial”

Figure 14. Core sub-network of $k = 16$ in ET-EMCM

Table 10. Top 10 core nodes of $k = 16$ in ET-EMCM, with higher Betweeness Centrality, Frequency and Group

<table>
<thead>
<tr>
<th>Node label</th>
<th>Frequency</th>
<th>Betweeness Centrality</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>model</td>
<td>72</td>
<td>5211.504666</td>
<td>1</td>
</tr>
<tr>
<td>design</td>
<td>45</td>
<td>2488.950239</td>
<td>0</td>
</tr>
<tr>
<td>flow</td>
<td>48</td>
<td>2151.301716</td>
<td>3</td>
</tr>
<tr>
<td>finite-element method</td>
<td>32</td>
<td>1341.887342</td>
<td>2</td>
</tr>
</tbody>
</table>
As shown in Figure 15, with 0.347 of modularity, there are 4 sub-groups as follow:

1) In Group 0, most keywords are related to Astronomy & Astrophysics, especially research of the sun, including structure and investigative techniques, such as “dynamics”, “magnetic-field”, “sun: activity”, “corona”, “sun: chromosphere”, “magnetohydrodynamics”, “dynamics-observatory sdo”, etc. Solar corona is an aura of plasma that surrounds the sun, as well as Solar Dynamics Observatory (SDO) is a satellite observing the sun launched by NASA. What’s more, as one of the Magnetohydrodynamics (MHD) systems, the sun is focused on by the study about magnetic properties of electrically conducting fluids.

2) Related to MHD, Group 1 has “magnetic reconnection”, “coronal mass ejections”, “flux”, “sun: magnetic topology”, “sun: surface magnetism”, “turbulence” and so on.

3) Compared with the two former subgroups, Group 3 and Group 2 are less homogeneous, and keywords of them belong to Optics, Polymer Science, Electrochemistry and Instruments & Instrumentation.

Figure 16 shows the core sub-network in IEAPO, with the K-core value set by 18, as well as Table 11 lists the top 10 nodes with frequency, betweeness centrality and modularity class. Compared with Figure 15, the core sub-network is similar to the mixture of Group 0 and Group 1. The keywords with high betweeness centrality over 300 are likely to concern with the study about MHD of the sun and its activity, such as coronal mass ejections and sunspot.
Figure 15. The communities of keywords in IEAPO

![Figure 15](image)

Figure 16. Core sub-network of $k = 18$ in IEAPO

![Figure 16](image)

Table 11. Top 10 core nodes of $k = 18$ in IEAPO, with higher Betweenness Centrality, Frequency and Group

<table>
<thead>
<tr>
<th>Node label</th>
<th>Frequency</th>
<th>Betweenness Centrality</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>model</td>
<td>32</td>
<td>1653.728205</td>
<td>1</td>
</tr>
<tr>
<td>dynamics</td>
<td>25</td>
<td>797.4006616</td>
<td>0</td>
</tr>
<tr>
<td>magnetic-field</td>
<td>20</td>
<td>449.1046422</td>
<td>0</td>
</tr>
<tr>
<td>sun: activity</td>
<td>25</td>
<td>405.0358103</td>
<td>0</td>
</tr>
<tr>
<td>plasma</td>
<td>11</td>
<td>330.1535197</td>
<td>0</td>
</tr>
<tr>
<td>flare</td>
<td>24</td>
<td>301.323679</td>
<td>1</td>
</tr>
<tr>
<td>emission</td>
<td>17</td>
<td>296.9724659</td>
<td>0</td>
</tr>
<tr>
<td>magnetic reconnection</td>
<td>17</td>
<td>214.048599</td>
<td>1</td>
</tr>
<tr>
<td>sun: uv radiation</td>
<td>28</td>
<td>203.239737</td>
<td>0</td>
</tr>
<tr>
<td>coronal mass ejections</td>
<td>34</td>
<td>185.7645547</td>
<td>1</td>
</tr>
</tbody>
</table>

Conclusion and limitations

The conclusion presented here are based on SCI-Expanded articles funded by LMT and published from 2009 to 2016. Although research funding is being more and more significant in academic research, it cannot meet all researches’ demand, and this is why funder has set regional laws to examine and verify application from research communities, selecting the most valuable project they thought. This behavior not only indirectly influences the development of S&T, but also reflects which science field or technology funding provider focuses on, especially private company. Throughout funded papers, it is possible to detect
funding agency’s intellectual structure, as well as scientific and technological frontier. This paper shows a feasible way to map the intellectual structure and find frontiers of S&T. The top 15 disciplines and three main clusters are completely conform to what area LMT engages in. Besides Aeronautics, Missiles and Fire Control, Rotary and Mission Systems, and Space Systems, the four main business areas, Lockheed has business in the energy and emerging technologies. And Sandia is home to a wide variety of research including computational biology, mathematics, materials science, alternative energy, psychology, MEMS, and cognitive science initiatives. In cluster analysis, the top 15 disciplines were divided into three main clusters, MSCP, ET-EMCM and IEAPO. Compared all of them, it is obvious that MSCP is much more homogeneous than the rest, and has a large number of papers. Materials Science, Chemistry and Nanoscience & Nanotechnology are the fields that develop crosswise recently, and are the academic areas LMT focused on. Besides, Energy & Fuels and Thermodynamics, Engineering and Mechanics, as well as Computer and Mathematics are also the areas LMT pays attention to, which are related to Lockheed’s business and Sandia’s research areas. With business of space systems, LMT shows interest in Astronomy & Astrophysics.

Some detailed science fields and technologies funded by LMT are founded. In social network analysis, thin film in nano-scale and its growth are the core study in MSCP, while ET-EMCM is specifically about scientific methods in these subjects, such as molecular dynamics simulation used in special material, and IEAPO is focus on the sun’s structure and activity, especially MHD. Except the major parts above, there are some special fields detected, such as nanowires applied in battery, microfluidics, phase as well as metamaterial used in artificial implements.

The method in this paper is feasible to detect information of S&T in LMT. However, there are four limitations to this paper. First, the datasets in this paper are collected in the WoS, other databases are ignored. Second, we just used corporation name as search string to collect datasets, ignoring funding program manipulated or supported by LMT. Thirdly, we just chose the main part of samples we collected and did not explored the whole datasets, which might contain less but more valuable information of LMT. Lastly, we viewed Sandia and Lockheed as a whole, which ignored difference between them. They should be separately further elaborated in future research.

Acknowledgments
This paper is supported by National Social Science Foundation in China (Grant No.16BTQ055).

References


Exploring the Similarity of Articles Co-cited at Different Levels

Giovanni Colavizza¹, Kevin W. Boyack², Nees Jan van Eck³ and Ludo Waltman³

¹ giovanni.colavizza@epfl.ch
Digital Humanities Laboratory, École Polytechnique Fédérale de Lausanne (CH)

² kboyack@mapofscience.com
SciTech Strategies, Inc. (USA)

³ {ecknjpvan, waltmanlr}@cwts.leidenuniv.nl
Centre for Science and Technology Studies, Leiden University (NL)

Abstract
We investigate the similarities of pairs of articles which are co-cited at the different co-citation levels of the journal, article, section, paragraph, sentence and bracket. Our results indicate that the following similarities rise monotonically as the co-citation level gets lower (from journal to bracket), in order of importance: textual similarity, intellectual overlap (shared references), author overlap (shared authors), proximity in publication time. While the main gain in similarity happens when moving from journal to article co-citation, all levels entail an increase in similarity, especially the paragraph and sentence/bracket levels. We compare results from three journals over the years 2010-2015, Cell, the European Journal of Operational Research and Physics Letters B, with consistent general outcomes.

Conference Topic
Citation and co-citation analysis, Natural Language Processing.

Introduction
The co-citation relation is used extensively in both information retrieval and bibliometrics. Applications include the identification of topically-related publications for search engines and clustering of publications to understand the structure of science. If two or more publications are co-cited by a third one, they are generally assumed to be related to some extent. Normally, this assumption is considered to be valid at a relatively coarse co-citation level, most often at the publication (e.g. article) level. In addition, recent work suggests that the similarity between co-cited publications might increase with increasing proximity of two publications within the full text of the citing publication, and that improvements in maps of science or document retrieval can be obtained by taking textual proximity into account. Indeed, it makes sense to assume that if two publications are co-cited in the same sentence or bracket in a publication, they typically will be in some way more related than two publications barely co-cited at the more general section or publication levels. Yet open questions remain and we know little about the ways in which related, co-cited publications are similar. This proof-of-principle study was designed to provide answers to these questions.

The recent increase in availability of full text data is beginning to impact the areas of research of bibliometrics and the retrieval of scientific publications by enabling the exploration of topics such as those listed above at a finer grained level. In this study we explore the implications of co-citations at different levels of granularity, and we do so by: a) considering different measures of similarity: textual similarity, reference overlap, shared authors, and publication time; b) comparing articles from three journals as representing different fields and possibly different citation practices and behaviors. The three journals considered in this study
are *Cell*, the *European Journal of Operational Research* and *Physics Letters B*. After discussing related literature in the following section, we define our dataset and methods, discuss empirical results, and then conclude with implications and suggestions for future research.

**Related Work**

Co-citation relations have been used in a variety of contexts. Originally introduced in 1973 independently by Small (1973) and Marshakova Skaikevich (1973), co-citation analysis has most often been used to study the structure of science from the perspective of cited publications. A pair of publications is considered to have been co-cited if they appear together in the reference lists of one or more citing publications. The more often a pair of publications has been co-cited together, the more related they are assumed to be. The notion of co-citation has been extended to cited authors (White and Griffith 1981) and journals as well (McCain 1991; Ding, Chowdhury & Foo 2000). While it was originally used only on very small data sets, in recent years it has been used to cluster sets of over two million publications (Boyack & Klavans 2014).

The increasing availability of the full text of scientific publications has enabled a rising interest in its use in bibliometrics and related areas of research (Jha et al. 2016). Recently, it was also shown using a set of nearly four million publications from Elsevier journals that full-text features can sensibly enhance the prediction of the future impact of a publication, and more broadly that “it is well worth the effort to obtain the full text of scientific articles and to exploit the power of natural language analysis” (McKeown et al. 2016).

Recent analysis has also been using sentences from full-text publications. For example, pairs of articles that are cited within the same sentence have been shown to be more similar than pairs of articles that are cited within the same article (Tran et al. 2009), which has obvious applications in information retrieval and document clustering. Liu et al. (2014), using several hundred thousand articles from PubMed Central, found that queries against citing sentences do very well at finding highly relevant older articles, but not for newer ones. Doslu and Bingol (2016) used citation contexts (text surrounding the reference marker) to improve the ranking of items in a retrieval context using data from CiteSeerX. Citation contexts have also been used for sentiment analysis (Small 2011) and for identifying biomedical discoveries (Small, Tseng and Patek 2017).

More closely related to our study are a number of works aimed at improving measures of publication relatedness using their full text. The majority of these have investigated co-citation proximity, while a couple are based on bibliographic coupling.

Nanba et al. (2000) were the first to test if co-citation proximity was related to increased textual similarity. Using a small set of citation areas (typically a couple of sentences) extracted by hand, they reported a rise in the textual similarity of two articles as their co-citation proximity increased. Elkiss et al. (2008) explored a set of nearly 2500 full-text articles from PubMed Central, finding that the cosine similarity of articles increased if they were co-cited at closer granularities, considering both their abstracts and body texts. Gipp and Beel (2009) suggested weighting the links between pairs of articles in a co-citation network according to their co-citation level, using full weight for the sentence level, half weight for the paragraph level, etc. They found that this weighting scheme performed twice as well at retrieving relevant publications than when not accounting for link weights. As mentioned above, Tran et al. (2009) found that pairs of articles that are co-cited within the same sentence
were more similar than pairs of articles that are co-cited only at the article level. Callahan et al. (2010) suggested using the structure of the citing publication to establish co-citation relations among publications, albeit without assessing their method with more than anecdotal evidence. Liu and Chen (2012) examined a large dataset of publications and resulting co-citation pairs from BioMed Central (BMC), analyzing the sentence, paragraph, section and article levels. They found that in general lower co-citation levels are correlated with higher co-citation frequency, supporting the use of co-citation level information if available, and suggesting that co-citations at the sentence level form the basic structure of the general co-citation network. Boyack et al. (2013) used co-citation proximity information, as mapped by character offset, in order to improve co-citation clustering of articles. Their results on nearly 300,000 full-text articles published in Elsevier journals in 2007 show that the textual coherence of resulting clusters can increase by up to 30%, at the price of a decrease in the size of clusters. Finally, in a work using Wikipedia entries rather than scientific articles, Schwarzer et al. (2016) found that citation proximity analysis improved recommendation quality as compared to a simple co-citation approach.

Despite the above efforts, and the evidence in support of the idea that lower co-citation levels entail an increased similarity between publications, not much is known on what kind of similarity is involved, nor if the effect is similar in different fields of research.

**Methods and Data**

**Definitions**

We consider a co-citation to be a relation established between a pair of publications which are both cited at a certain level within scientific text. In this study we consider articles as publications, our hierarchy therefore starts at the journal level, where articles are published. The levels at which a pair of publications can be co-cited are, from high to low:

1. **Journal**: publications co-cited within the same journal. In our analysis, we consider articles as co-cited within the same journal if they are cited in the journal in the same year.
2. **Article**: publications co-cited within the same article.
3. **Section**: publications co-cited within the same section – a logical unit of the publication identified by some header – in an article.
4. **Paragraph**: publications co-cited within the same paragraph in an article. Paragraphs are usually identified with some layout expedient such as indentation or interlinear space.
5. **Sentence**: publications co-cited within the same sentence in an article.
6. **Bracket**: publications co-cited at the same location in an article. These are often delineated with brackets or parentheses.

Note that a pair of publications co-cited at a given level is also to be considered as co-cited at any higher level. For instance, if two publications are co-cited at the paragraph level, then they are also considered as co-cited at the section, article and journal levels. In the literature co-citation proximity often, but not always, entails considering the textual distance between co-cited publications, such as character offset (e.g. Boyack et al. 2013). In our case, we do not take this approach. Instead, we use meta (e.g. publication), logical (e.g. section) or syntactic (e.g. sentence) textual units to identify different co-citation levels. The two approaches are related but not identical. We therefore refer to co-citation levels instead of co-citation proximity. We order co-citation levels from high (journal) to low (bracket). In this study we only consider journal articles as co-cited publications for analysis.
Data

Our analysis considers three Elsevier journals: Cell, the European Journal of Operational Research (EJOR), and Physics Letters B (PLB). Our analysis includes journals from three very different disciplines, possibly with quite different citation practices. Cell is a prominent life sciences journal. EJOR can be seen as a journal at the interface between the social sciences, computer science, and mathematics. PLB is a physical sciences journal. The three journals have also been chosen because they are all quite large and because they all have at least a moderately high citation impact. We analyze co-citations in the three above-mentioned journals in the six-year time period 2010-2015. We have chosen to consider a six-year time period to make sure that we have a sufficiently large number of data points per journal and to guarantee the reliability of our analysis.

The dataset of pairs of co-cited articles was constructed as follows. First, we used every article published from 2010 to 2015 in the three journals under consideration classified as either “full research article” or “short communication”. Review articles were excluded since their specific co-citation patterns may be different from articles describing original research. Second, we restricted the cited articles (references) to those which are published in Elsevier journals. Although this did reduce the number of potential co-cited pairs to be studied, this restriction facilitated our study in that it allowed easy access to the abstracts and cited references of cited articles. It is important to note that, by considering only cited articles published by Elsevier, we still retained between one third and half of all cited articles for analysis (Table 1).

Given this selection, we defined six sets of co-cited article pairs for every journal, corresponding to the six co-citation levels defined above. At the article level and the lower levels, all co-cited article pairs were used. Co-cited article pairs at the journal level were sampled from the large number of possible pairs. For each pair of articles co-cited within the same article, the most detailed level at which it is co-cited was identified, and the pair was counted for that level and all higher levels up to the article level. The frequency of co-citations was not considered. Consider a pair of articles that have been co-cited multiple times at the article level (because at least one of the articles has more than one in-text mention). Suppose that one of the co-citations is at the bracket level. The pair of articles will then be considered as co-cited at the bracket level and consequently also at all higher levels.

The hierarchy of the number of pairs per level is shown in Table 1; for instance, Cell has 9,392 pairs at the sentence level of which 5,585 are also at the bracket level. For the journal level, we created a large set of unique pairs of co-cited articles for each journal by sampling with replacement from the available pool of articles that have been cited by articles from the given journal in the same year. The number of pairs maintained at the journal level is somewhat higher than the numbers at the article level. Characterization of the dataset at all levels for each journal is given in Table 1.

Table 1. Basic statistics of the dataset for each journal. All co-cited pairs counts consider only co-citations of articles that are present in the Elsevier full-text dataset.

<table>
<thead>
<tr>
<th></th>
<th>Cell</th>
<th>EJOR</th>
<th>PLB</th>
</tr>
</thead>
<tbody>
<tr>
<td># Citing articles</td>
<td>2,038</td>
<td>3,463</td>
<td>4,887</td>
</tr>
<tr>
<td># References (total)</td>
<td>109,377</td>
<td>111,941</td>
<td>166,841</td>
</tr>
<tr>
<td># References (Elsevier)</td>
<td>46,269</td>
<td>52,789</td>
<td>60,453</td>
</tr>
<tr>
<td>% References (Elsevier)</td>
<td>42%</td>
<td>47%</td>
<td>36%</td>
</tr>
<tr>
<td># Co-cited pairs – Journal level</td>
<td>245,703</td>
<td>277,150</td>
<td>283,801</td>
</tr>
<tr>
<td># Co-cited pairs – Article level</td>
<td>177,123</td>
<td>216,846</td>
<td>159,111</td>
</tr>
</tbody>
</table>
Measures
The similarity between two publications can be conceptualized in many different ways. Here we focus on four main axes: textual similarity, intellectual overlap, author overlap and time distance.

To establish the textual similarity between two publications in a pair we considered their titles and abstracts and use the BM25 measure, widely adopted to rank documents for information retrieval given a textual query (Spark Jones et al. 2000) and more recently also used to cluster publications using their titles and abstracts (Boyack et al. 2011). Each publication (title and abstract) was reduced to lowercase and split into tokens for calculation of textual similarity (eliminating punctuation and then tokens of just one alphanumeric character). Given a publication \( q \) and another publication \( d \), BM25 similarity was calculated as

\[
s(q,d) = \sum_{i=1}^{n} IDF_i \frac{n_i(k_1+1)}{n_i+k_1(1-b+0.75|D|/|D|)},
\]

where \( n \) denotes the number of unique tokens in \( q \), \( n_i \) equals the frequency of token \( i \) in document \( d \), and \( n_i = 0 \) for tokens that are in \( q \) but not in \( d \). \( k_1 \) and \( b \) have been set to the commonly used values of 2 and 0.75 respectively. \(|D|\) denotes the length of document \( d \), in number of tokens. \(|D|\) denotes the average length of all documents in the collection. The IDF value for every unique token in the collection was calculated as

\[
IDF_i = \log \left( \frac{N-d_i+0.5}{d_i+0.5} \right),
\]

where \( N \) denotes the total number of documents in the collection and \( d_i \) denotes the number of documents containing token \( i \). IDF scores strictly below zero were discarded to filter out very commonly occurring tokens. BM25 is not a symmetric measure. We obtained a symmetric measure for the similarity of documents \( q \) and \( d \) as follows:

\[
\frac{s(q,d) + s(d,q)}{2}.
\]

A BM25 textual similarity was calculated for every article pair listed in Table 1. For each journal, we then divided by the maximum BM25 value among pairs from that journal in order for the measure to range between 0 and 1. We note that after applying this normalization BM25 values cannot be directly compared between journals.

We define the intellectual overlap to be the proportion of references that a pair of publications have in common (alternatively known as bibliographic coupling):

\[
\frac{N_{qd}}{\min(N_q,N_d)},
\]

where \( N_{qd} \) denotes the number of overlapping references of publications \( q \) and \( d \) and \( N_q \) and \( N_d \) denote the total number of references in publications \( q \) and \( d \) respectively. The intellectual
overlap equals one if all references of the publication with the shorter reference list are also cited by the other publication.

We define the author overlap as the proportion of authors that a pair of publications have in common. Analogous to the intellectual overlap measure we used the minimum number of authors of the two publications as the normalization basis (similar to Eq. (4)). If all authors of the publication with fewer authors are also authors of the other publication, the author overlap equals one. Matching of authors was done using the following process. Starting with a string of text for the surname of an author and one for the full name of the author, we converted both strings to lowercase. We then compared all possible combinations of author mentions. We considered two mentions to refer to the same author if the surname was identical and all components of the full name of the shorter name string had a match in the longer name string. For example, “John J. Abrams” will match with “J. Abrams” but not with J. M. Abrams”. Given that we operate within the context of a specific journal and that the procedure is fairly conservative, we deem the results as sufficiently accurate.

Lastly, we define the time distance as the number of months between the publication dates of a pair of co-cited publications. We used the publication month data present in the Elsevier full text data.

Results

Distributions

We start our analysis by considering the distributions of the similarity measures of pairs of articles co-cited at different levels for each journal under consideration. Figure 1 shows the results for textual similarity calculated using the BM25 measure. Clearly, textual similarity monotonically increases as the co-citation level goes down. The means are located above the medians, suggesting that the distributions of the similarities are right skewed. A closer inspection of the distributions of the similarities (not shown here) reveals that the distributions are fairly normal, but indeed with a slight skew on the right side. For all three journals the main increase in similarity is obtained by moving from the journal to the article co-citation level. The section level adds little afterwards. Behaviors then differ: whilst article pairs co-cited in EJOR and PLB gradually increase in similarity, this increase is more pronounced for Cell.
We consider next the intellectual overlap of pairs of articles with results shown in Figure 2. Intellectual overlap presents a generally more skewed distribution, and an even larger gain in similarity takes place from the journal level, where similarity is essentially zero, to the article level. Here too, the similarity rises monotonically as we move down in the co-citation levels. PLB has the highest intellectual overlap in co-cited articles. The rise in intellectual overlap is also more rapid for PLB, ending up at the bracket level with an average number of nearly one in every four references being shared. Pairs of articles co-cited in Cell, as well, present a rapidly growing overlap as the level lowers. This effect is less pronounced in EJOR.

With respect to author overlap, results are given in Figure 3. Its distributions are very skewed, as at any level most pairs of co-cited articles do not share authors at all. Nevertheless, we see that author overlap in the case of PLB is stronger than in the case of the other two journals. We note that articles published in and cited by PLB have a more skewed distribution of the number of authors than those of Cell and EJOR. Physics is, in general, known to have a larger number of kilo-authored papers than other fields. Cell, on the other hand, has a higher median number of authors per article, as shown in Error! Reference source not found., which might lower its relative author overlap.

<table>
<thead>
<tr>
<th>Table 2. Mean and median number of authors of the citing and cited articles.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell</td>
</tr>
<tr>
<td>Mean (median) number of authors of citing articles</td>
</tr>
<tr>
<td>Mean (median) number of authors of cited articles (Elsevier only).</td>
</tr>
</tbody>
</table>
Lastly, in Figure 4 we consider the distance of the publication time for pairs of co-cited articles. *Cell* and *EJOR* in general refer to research more proximal in time as evidenced by their lower time distances at the journal and article levels. However, this difference nearly disappears with increasing proximity. As we move down in the co-citation levels, co-cited articles become increasingly similar in terms of their publication time, ending up at a median difference in publication time of about three years for each of the three journals at the bracket level.
Figure 4. Mean and median publication time distance (in months) of article pairs co-cited at different levels. Note that statistics are calculated at each discrete level, lines are only meant for readability.

Taken together these first impressions highlight how in general, and over four different dimensions (text, references, authors and time), co-cited articles that are located at lower levels tend to be more similar than those that are located at higher levels. Results confirm the intuitive hypothesis that, as the co-citation level lowers, research being co-cited is increasingly more related. The agreement of all similarities supports the idea that, especially at lower co-citation levels, authors tend to refer to coherent pieces of research that are similar in content, often similar in terms of authors involved, and that have been published around the same time.

Cumulative distributions
We now take another view by considering cumulative distributions. For each journal, each similarity measure and each proximity level, we inspect the proportion of pairs of co-cited articles with a similarity value equal to or below a certain threshold value. All results are shown in Figure 5. Starting with textual similarity, we can immediately appreciate how gains in similarity are not negligible at most levels. This is especially the case for pairs of articles co-cited in Cell, which seem to increase their similarity in three steps beyond the journal level: article/section, paragraph and sentence/bracket. For EJOR and PLB these steps are less well defined. The main increase takes place between the section and paragraph levels for these journals.

The cumulative distribution of the intellectual overlap follows slightly different patterns. The similarity of pairs of co-cited articles according to intellectual overlap is indeed more pronounced in PLB than in Cell and EJOR. Three major increases in similarity can be observed for Cell and PLB, namely at the article/section, paragraph and sentence/bracket levels, whilst for EJOR the only significant increase takes place between the section and paragraph levels.
Lastly, we consider the same distributions for the author overlap and the publication time distance. With respect to the author overlap, we notice how pairs of co-cited articles which share authors are relatively rare at all levels, as was indeed evident already from Figure 3, and the importance of this aspect for PLB is particularly remarkable. In a similar way to what we discussed for the intellectual overlap, the author overlap seems to follow a three step pattern in Cell and PLB. The distributions for the publication time distance are more compact, except for PLB, where gains in time distance moving to lower levels are more important.

All journals share an increase in similarity as we move to lower co-citation levels. The main transition happens between the journal and article co-citation levels. Cell and EJOR article pairs become relatively more similar with respect to textual similarity and time, PLB pairs with respect to their intellectual and author overlaps. A further distinction appears when we consider relative similarity gains at different levels. While Cell and PLB in general show noticeable similarity increases at three stages (article/section, paragraph and sentence/bracket), EJOR is more gradual and presents only two clear stages (article/section and below).
Conclusions
In this article we explored the similarity of article pairs which are co-cited at different levels. We considered three journals from different disciplines: Cell, the European Journal of Operational Research (EJOR) and Physical Letters B (PLB). Our results indicate that the similarity of pairs of articles increases monotonically with their co-citation level. In other words, the lower the level at which two articles are co-cited, the more similar the articles will be on average. We used different measures of similarity between pairs of co-cited articles: textual similarity, intellectual overlap, author overlap and distance in time. These measures all increase as the co-citation level lowers.

Our results provide a solid confirmation of the idea that there is value in using co-citation proximity information in order to capture stronger links between co-cited articles. This result has implications for, among others, science mapping and the retrieval of scientific literature. However, this study is not without limitations. In particular, it explores only three journals and thus cannot truly represent co-citation behaviors by discipline or field. Additional studies with larger data sets are needed to explore field-based differences. In addition, this study explores similarity measures independently without considering their interactions. Other similarity measures could be explored as well, for example by considering the frequency of in-text citations.

Acknowledgments
Colavizza benefits from a Swiss National Fund grant, number P1ELP2_168489.

References


Is the Relationship between the Impact Factor and Papers’ Citations Really Weakening?

Yue Mingliang¹ Ma Tingcan²

¹ yueml@whlih.ac.cn ² matc@whlih.ac.cn
Wuhan Documentation and Information Center, Chinese Academy of Sciences, Wuhan (China)

Abstract
Recently, certain researches declared a weakening relationship between impact factor (IF) and papers’ citations due to the increasingly convenient access to individual papers in the digital age. In this paper, we study the publication distribution of the most cited papers on the most cited journals (i.e., journals with highest IFs) based on three pieces of data, i.e., Science Citation Index Expanded (SCI-EXPANDED), Essential Science Indicators (ESI) and Journal Citation Reports (JCR). The results show that either from the overall or research field perspective, there is no obvious clue to indicate the weakening relationship between IF and papers’ citations.

Conference Topic
The relationship and development of five metrics science concepts, that is, Bibliometrics, Informetrics, Scientometrics, Webometrics and Knowledgometrics.

Introduction
Ever since it was proposed, the discussion around Impact Factor (IF) has never stopped (Garfield, 2006; Russell, 2002). The debates always focus on whether IF is an effective way for research evaluation and arouse many variants of IF, e.g., the source normalized impact per paper (Moed, 2010), the Eigenfactor (West & Bergstrom, 2010), and the SCImago Journal Rank (González, 2009), etc. In all those works, IF is treated as an important factor that can influence the potential citations a paper can receive (Waltman, 2016). However, most recently, certain researches declared a weakening relationship between IF and papers’ citations due to the increasingly convenient access to individual papers in the digital age (Lozano, 2012). Whereas many others insisted that IF is still one of the most authoritative scientometric indicators for assessing the academic impact of journals and is influential to papers’ citations.

Focusing on this debate, in this paper, we study the publication distribution of the most cited papers on the most cited journals (i.e., journals with highest IFs) to analyze the relationship between IF and papers’ citations. Rather than ranking papers and journals with self-calculated citations and IFs (Lozano, 2012), we use citations and IFs published by Web of Science (WoS) and Incites. Further, we use IFs for journal ranking and Essential Science Indicators (ESI) baseline for top paper ranking. We believe this way is more objective and effective since WoS/Incites and ESI data is most widely accepted and used by scholars (Archambault, 2008; Meho, 2007). That means, if IF does have certain relationship with paper citation, it is WoS IF (rather than self-calculated IF) that affects scholars’ impression to a paper and the willing of citing it. The results show that either from the overall or research field perspective, there is no obvious clue to indicate the weakening relationship between IF and papers’ citations.

Method
We made use of three pieces of data in our analysis, i.e., Science Citation Index Expanded (SCI-EXPANDED), Essential Science Indicators (ESI) and Journal Citation Reports (JCR). We first downloaded the bibliography data of papers covering all areas of natural sciences from SCI-EXPANDED in between 2005-2014, and mapped every paper to the corresponding
research field according to the paper’s publishing journal based on the ESI category. Then we associated every paper with a top-level (e.g., Top 1%) based on the paper’s citation counts and the ESI citation baseline by considering the paper’s research field and publishing year, and used the Top 10% papers for statistics, involving 1.44 million papers and 20 research fields. Fig. 1 and Fig. 2 shows the number of papers published in each year and each research field respectively.

![Figure 1 Paper counts vs. Publishing years](image1)

Further, for every research field, we ranked its belonging journals according to their Impact Factors (IF) in the decrease order year by year, and calculated the top-level (e.g. Top 5% journals) for every journal in every year. Finally, we calculated the yearly publication distributions of the Top \(m\)% papers on the Top \(n\)% journals to assess the overall relationship between IF and papers’ citations and its variation trend, for \(m=0.1, 1, 10\) and for \(n=1, 5, 10\), where Top \(x\)% papers means papers with citation ranked in the Top \(x\)% and Top \(x\)% journals means journals with IF ranked in the Top \(x\)%. Moreover, we also carried out statistics in a more detailed perspective, i.e., research field, to test whether the overall trend holds in most research fields.
Results

Overall results

Fig. 3 to Fig. 5 show the percentages of the Top 0.1%, 1% and 10% papers published in the Top 1%, 5% and 10% journals, for all research fields and all countries (regions). From the figures, the following conclusion can easily be drawn: 1) in almost all the years, more than 80% of the Top 0.1%, 1% and 10% papers come from the Top 1% journals, and in average 88.9%, 87.2% and 86.3% of the Top 0.1%, 1% and 10% papers come from the Top 1% journals; 2) despite of small variations, the percentages of the Top 0.1%, 1% and 10% papers published in the Top 5% (10%) journals are all larger than 97% (99%); 3) none of the Figures reflects an obvious declining of the percentages (of top papers published in top journals). That is, in general best papers were still published in best journals, and there is no evidence to indicate the weakening relationship between IF and papers’ citations.

Figure 3 Percentages of the Top 0.1%, 1% and 10% papers published in the Top 1% Journals

Figure 4 Percentages of the Top 0.1%, 1% and 10% papers published in the Top 5% Journals
Field perspective

Here we aim to see whether the same relationship (among the top papers and the top journals) holds from the perspective of different research fields. To fully understand the relationship among the top papers and the top journals, here we focus on the Top 0.1% papers for illustration. In addition, the full statistic results including all the research fields and all the Top levels of papers and journals can be found in our supplementary materials†. From the results, we see that first, for the Top 1% journals, a) the percentages of the top papers published in the top journals are relatively low in certain research fields. As shown in Fig. 6, these fields include Material Science, Microbiology, Molecular Biology & Genetics, Physics and Psychiatry/Psychology. However, b) for all the remaining 15 research fields, the percentages of the top papers published in the top journals are approximately the same as the overall situation. As shown in Fig. 7, most of the percentages are larger than 80%, and in average 94.0% of the Top 0.1% papers come from the Top 1% journals.

† The supplementary materials are interactive dashboards formatted as PDF files, please refer to either of the following URLs: Google Drive: https://drive.google.com/file/d/0BwdzLr3JJ8qcZlJTMXJ3aWd3ZDQ/view?usp=sharing OR Baidu Cloud: http://pan.baidu.com/s/1mihsBO0 (Password for Access: zsni)
Second, for the Top 5% and 10% journals, in almost all the research fields and years, more than 90% (95%) of the Top 0.1% papers come from the Top 5% (10%) journals, as shown in Fig. 8 (for Top 5% journals) and 9 (for Top 10% journals).
From the above analysis, it can be seen that same as the overall situation, there is no obvious evidence to indicate the weakening relationship between IF and papers’ citations.

**Conclusion**

In this paper, we studied the relationship between IF and papers’ citations. The results show that either from the overall or research field perspective, there is no obvious clue to indicate the weakening relationship between IF and papers’ citations. That means IF still acts as a very important indicator for the quality of academic journals and can influence the potential citation a research paper can receive. We believe that at least three reasons contribute to this fact. First, higher IF can still attract papers with higher quality for publication. Then, higher IF journals may have scholars with higher academic attainments for peer review. Hence only papers with higher quality may appear on the journals. Those two facts guarantee the papers’ potential academic influence. Third, it is exactly because of the increasingly convenient access to individual papers in the digital age, that researchers can easily acquire large amounts of research results. Finding more appropriate references with higher quality becomes an urgent need. And it seems that currently there is no other way more effective and efficient than selecting based on IF and paper citation.

**Acknowledgments**

This work is supported by the National Natural Science Foundation of China under grant (No. 7160325); the Young Talent-Field Frontier Project of Wuhan Documentation and Information Center, Chinese Academy of Sciences.

**References**


UK ethnic minority cancer researchers: their origins, destinations and sex

Mursheda Begum¹ Philip Roe¹,² Richard Webber³ Grant Lewison¹,²

mursheda.begum@kcl.ac.uk, philip.roe@kcl.ac.uk, grant.lewison@kcl.ac.uk
¹ King’s College London, Department of Cancer Studies, London SE1 9RT (UK)

philip@evaluametrics.co.uk, grantlewison@aol.co.uk
² Evaluametrics Ltd, 157 Verulam Road, St Albans AL3 4DW (UK)

richardwebber@originsinfo.eu
³ Bisham Gardens, London N6 6DJ (UK)

Abstract
We identified the names of all authors of Web of Science papers on cancer in 2009-11 and in 2014-16 with a UK address, and listed those in each postcode area (the first one or two letters of UK postcodes). Their ethnicity or national origins were determined from their names with the OriginsInfo database into seven main groups; African, Asian excluding South Asian, South Asian, Western European, Eastern European, Islamic World, and the UK/Ireland. We were thus able to determine the geographical distribution within the UK of members of the six immigrant groups in terms of their fractional contributions to cancer research. The percentage of research contributions from scientists with non-British (or Irish) names rose from 37% in 2009-11 to 39.4% in 2014-16. It was higher in London, and rose from 47% to 51%. The largest ethnic contributions were from Europe: in 2014-16 the percentages were 15.4% from Western Europe and 3.7% from eastern Europe. There were 8.1% of contributions from South Asia, and 5.6% from the rest of Asia; the largest were in four areas of London, and in Cambridge and Oxford. Female participation is still less than that of males, but it is increasing by about 2% per year.

Conference Topics
Country-level studies; mapping and visualization; participation in science; science policy

Introduction
This study began with an investigation of the role of immigrant scientists in UK science, and in particular with their presence at the higher levels - Nobel laureates, Fellows of the Royal Society and Fellows of the Academy of Medical Sciences. We showed that immigrants had played a substantial role at these higher levels, and that any restrictions that might be imposed on scientists from the European Union (EU) coming to the UK when it was no longer a Member State would have damaging consequences (Begum et al., 2016). The investigation also looked at the presence of scientists with foreign names among UK cancer researchers in 2009-11, and showed that about three quarters of such groups who authored papers recorded in the Web of Science (WoS, © Clarivate Analytics) contained at least one member with a foreign name, either continental European or from the Rest of the World (which would not be affected directly if the UK left the EU).

The present study builds on this, and examines the UK locations where the ethnic minority cancer researchers are working, the world regions from which they come, and (for most of them) their sex. It also examines two separate cohorts: those publishing papers in 2009-11, and a later one from 2014-16. It is intended to provide further evidence of the part played by immigrant researchers, and to show where they are most productive. It will also provide information on the sex balance of cancer researchers, both those with British names and those from the different ethnic groups or with different national origins.

There is a paucity of literature on the ethnicity or national origins of researchers. However, it has been shown that immigrants can make substantial contributions to the scientific output of their host countries in different disciplines. For example, Webster (2004) showed the
importance of different ethnic groups to UK science in several major fields. There are more Indian cancer researchers in Canada and the USA than in India itself (Basu, Roe & Lewison, 2012). In lung cancer research, there has also been much international migration (Lewison et al., 2016), and in Canada and the Nordic countries many of the researchers were from ethnic minorities.

In the UK, various organisations collect personal information on ethnicity as part of the concern to improve opportunities for minorities. The forms offer a number of options, but cannot be comprehensive or deal adequately with people of mixed race or ethnicity, and some respondents decline to answer. Although names do not necessarily provide an exact indicator of nationality, studies have consistently demonstrated (Webber, 2007; Mateos et al., 2007; Webber 2010) a very close correspondence between a person’s given and family names, taken in combination, and the cultural background of their forebears. They can also indicate a degree of cultural assimilation when people have a given name indicative of a different ethnicity than that of their surname, suggesting that they are second-generation rather than first-generation arrivals. This is explored in a little more detail in the Discussion section.

Methodology

The selection of cancer research papers.
Cancer papers were identified using a proprietary filter (Lewison, 2011) developed in consultation with Cancer Research UK and updated with Escuela Andaluza de Salud Pública; our Spanish partners on a previous EU mapping project. The filter comprehensively listed specialist cancer journals and title words that covered several cancer manifestations, genes implicated in one’s chance of developing or avoiding cancer, and drugs exclusively used for cancer treatment. It had a strong precision (0.95) and recall (0.98) and was applied to Clarivate Group’s Web of Science (WoS). The WoS literature search was limited to articles and reviews from both the Science Citation Index and Social Sciences Citation Index, which had at least one address in the UK, and were published in two specific time periods; 2009-11 and 2014-16. These papers were downloaded as text files, 500 at a time, and were then converted into a single Excel spreadsheet for each time period using a macro developed for WoS papers by Evaluametrics Ltd.

Since 2009, the WoS has tagged the authors of papers with their individual addresses in a column labelled “C1”, and these data were used to distinguish authors with a UK address from ones working abroad. Another macro listed the names of each UK author, with her or his postcode area (see below) and their total fractional contribution, based on the numbers of authors on each paper. Authors with an address in the UK and one elsewhere would have had their contribution divided between their different addresses, and similarly if they worked in two (or more) UK postcode areas.

Origins: Ethnicity

The list of cancer researchers with UK addresses with their given and family names was sent to OriginsInfo Ltd for classification by country of origin/ethnicity and sex using the OriginsInfo database. This is a large proprietary database of approximately four million surnames and one million given names. Ethnicity can be determined in 99.5% of cases when both names are present, and with 96% accuracy on the basis of surnames for higher-level ethnic categories (e.g. West African) and at 93% at lower-levels (e.g. Nigerian and Ghanaian). The ethnic categorisation of researchers was very detailed, sometimes distinguishing between distinct regions of the same country (e.g. Indian categories included Gujarati, Punjabi, Hindi, Bengali, Tamil, Sikh, etc.). Subsequently, this information was condensed to 11 main world regions identified by three-letter codes listed in Table 1. Some
“authors” were collaborative groups, and so were not assigned to a world region. Some authors were not assigned to a world region by the software, and the names were inspected so as to categorize some of them where it appeared clear which region they were from. Of the 78,101 names, 69 were groups and only 392 could not be assigned to a region (0.5%).

Table 1: Details of the 11 world regions by which cancer researchers were categorised.

<table>
<thead>
<tr>
<th>Code</th>
<th>World Region</th>
<th>Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFR</td>
<td>Africa</td>
<td>All countries in Africa excluding Egypt.</td>
</tr>
<tr>
<td>ASI</td>
<td>Asia</td>
<td>All countries in Asia excluding those from ISL and SAS.</td>
</tr>
<tr>
<td>AUS</td>
<td>Australasia and Oceania</td>
<td>All countries in Australasia.</td>
</tr>
<tr>
<td>EEU</td>
<td>Eastern Europe</td>
<td>The &quot;new&quot; Member States of the European Union (Bulgaria, Croatia, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia, Slovenia) plus Albania, Armenia, Azerbaijan, Belarus, Bosnia &amp; Herzegovina, Georgia, Kosovo, Macedonia, Moldova, Montenegro, Russia, Serbia, Ukraine</td>
</tr>
<tr>
<td>EUR</td>
<td>Western Europe</td>
<td>The &quot;old&quot; Member States of the European Union (Austria, Belgium, Denmark, Finland, France, Germany, Greece, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden) plus Cyprus, Iceland, Malta, Norway and Switzerland, minus the UK and Ireland.</td>
</tr>
<tr>
<td>ISL</td>
<td>Islamic World</td>
<td>Bahrain, Egypt, Iran, Jordan, Kuwait, Lebanon, Oman, Qatar, Saudi Arabia, Turkey, Yemen and United Arab Emirates</td>
</tr>
<tr>
<td>LAT</td>
<td>Latin America</td>
<td>South and Central America, and Mexico.</td>
</tr>
<tr>
<td>NAM</td>
<td>North America</td>
<td>Canada and the USA.</td>
</tr>
<tr>
<td>SAS</td>
<td>South Asia</td>
<td>Bangladesh, India, Pakistan, and Sri Lanka.</td>
</tr>
<tr>
<td>UKI</td>
<td>UK/Ireland</td>
<td>United Kingdom and Ireland.</td>
</tr>
<tr>
<td>WIN</td>
<td>West Indies</td>
<td>Greater and Lesser Antilles, and Lucayan Archipelago.</td>
</tr>
</tbody>
</table>

Destinations: Postcodes

Geographical distributions within the UK are conveniently described by postcodes, which form part of the address. The UK postcode operates at four levels (Shepherd et al., 1992):

- the postcode area, consisting of one or two letters (e.g., SE, G);
- the district, consisting of these plus one or two numbers, (e.g., SE1)
- the sector, consisting of these, plus a space and a single number (e.g., SE1 6);
- the full postcode, consisting of these plus two letters (e.g., SE1 6RT).

The full postcode normally defines a street, or part thereof, or a large building such as a hospital, but in our analysis we only used the first part, the postcode area, of which there are 124 for the UK, including the Isle of Man (IM) and the Channel Islands (Guernsey GY and Jersey JE). London has eight postcode areas: E, EC, N, NW, SE, SW, W and WC. The five major university hospital groups are mostly associated with one of these: Imperial with W, King's College with SE, Queen Mary (Bart's and the Royal London) with E and EC, St George's with SW, and University College with WC and NW (Royal Free).

An address macro was applied to the C1 column of the spreadsheet to identify all UK addresses with their postcode areas. This created two worksheets. One listed each author with an identifiable UK postcode in the C1 column together with their fractional
contributions to cancer research, and a second with any addresses without a postcode present or not in the address thesaurus. These needed to be manually searched and coded before the macro was subsequently re-run. This process characterised almost all UK addresses with their postcode areas, but some turned out not to be UK (e.g., they were in London, Ontario) and were coded “XX” or were too vague (e.g., University of London, London) and were left blank, or simply could not be found from a search on the Web. Altogether, we were able to find the postcode areas for all but 0.04% of the UK addresses.

Determination of the authors’ sex

This was the most difficult part of the exercise and we undertook the sexing of individual authors in five stages. The first was through the OriginsInfo database, which was able to assign sex to many of those with given names. In 2009-11, 33% of authors only had one or more initials, but by 2014-16, this proportion had reduced to 25%; this probably reflects the changing policy of journals and the desire of authors to distinguish themselves from others with the same surnames and initials.

The second stage involved the analysis of another large database, that of over 300,000 UK-registered doctors, which lists their given names and their sex. We tabulated all the different given names, with the numbers who were male and female. Of 51,865 different given names, 2,830 could be classed as male and 2,145 as female, on the basis of there being at least five people with the name of one sex only, or at least ten times as many of one sex as of the other. For example, there were 58 Ranjits who were male but only four female, so this name was treated as male, but there were 36 Padmas who were female but only two male, so this name was treated as female.

The third stage was to send lists of given names of people from particular regions or individual countries to people from these countries whom we knew (see Acknowledgement), who might be expected to know which given names were male and which female; if they were not sure or the names could be of either sex, they were not sexed.

The fourth stage involved the use of a special macro (written by PR) that examined names with only initials, and sought other authors with the same surnames and given names with the same initial letter. If a match was found, the possible matching given names were added. We then checked the sexes of these, and if and only if they were all either male or female (not blank), then we assigned a sex to the name with only initials. Thus Abel, P matched with both Abel, Paul and Abel, Peter, both of whom were male, so Abel, P was male. But Adams, C matched with both Adams, Christopher and Adams, Caroline, so could not be sexed. Some of these names could, however, be sexed in a fifth stage if the person with just an initial was working in the same postcode area as a person with a sexed given name and the others were working elsewhere.

Results

There were totals of 14,373 papers for 2009-11 and 17,107 for 2014-16, and there were 34,347 names and 43,754 names, respectively. Some of these were clearly the same people working in different postcode areas, and there were also name variants (either initials or given name, or sometimes both, e.g., Abdelrahman, Mostafa H.). The individual contributions from authors from each world region and each sex to each postcode area were listed by means of pivot tables for the two time periods, so that the regional contributions could be seen, and also the distribution by sex.
 Origins: ethnicity or national background

Table 2 shows the cancer research fractional contributions, and their percentages of the total, for the different ethnic groups in 2009-11 and in 2014-16.

Table 2. Ethnic composition of UK cancer researchers, 2009-11 and 2014-16. See Table 1 for regional codes.

<table>
<thead>
<tr>
<th>Region</th>
<th>2009-11</th>
<th>2014-16</th>
<th>% change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Contribution</td>
<td>Percent</td>
<td>Contribution</td>
</tr>
<tr>
<td>UKI</td>
<td>5925</td>
<td>63.0</td>
<td>5875</td>
</tr>
<tr>
<td>EUR</td>
<td>1343</td>
<td>14.3</td>
<td>1491</td>
</tr>
<tr>
<td>SAS</td>
<td>744</td>
<td>7.9</td>
<td>781</td>
</tr>
<tr>
<td>ASI</td>
<td>483</td>
<td>5.1</td>
<td>538</td>
</tr>
<tr>
<td>ISL</td>
<td>369</td>
<td>3.9</td>
<td>379</td>
</tr>
<tr>
<td>EEU</td>
<td>268</td>
<td>2.9</td>
<td>356</td>
</tr>
<tr>
<td>AFR</td>
<td>111</td>
<td>1.2</td>
<td>136</td>
</tr>
<tr>
<td>LAT</td>
<td>61</td>
<td>0.65</td>
<td>70</td>
</tr>
<tr>
<td>AUS</td>
<td>21</td>
<td>0.23</td>
<td>15</td>
</tr>
<tr>
<td>NAM</td>
<td>13</td>
<td>0.14</td>
<td>6</td>
</tr>
<tr>
<td>WIN</td>
<td>7</td>
<td>0.08</td>
<td>8</td>
</tr>
<tr>
<td>Unknown</td>
<td>49</td>
<td>0.52</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>9398</td>
<td></td>
<td>9690</td>
</tr>
</tbody>
</table>

The researchers with UK or Irish names form the largest contingent, of course, but "ethnic minorities" now make up about 40% of the total. The largest groups are the West Europeans and South Asians, but the East Europeans are growing much faster (nearly 29% increase between the two periods). The Africans and Latin Americans are also growing, but from much smaller bases. Figures 1 and 2, below, show the countries involved in the West and East European contingents, and from the Asian ones, for the whole six-year period. The digraph ISO2 codes used in the pie charts are identified in Table 3.

Figure 1. Contributions to UK cancer research in 2009-11 and 2014-16 combined by researchers from the "old" European Union (left) and from the "new" EU and other Eastern European countries (right); fractional counts of papers. Areas of pie slices are proportional to contributions.
Cancer research in the UK is concentrated in the "golden triangle" of Cambridge, London and Oxford, as Table 4 clearly shows. The 10 postcode areas for these cities accounted for 38% of the total UK contribution in the six years (data for London N not shown, but equal to 0.12%).
Table 4. Contributions to UK cancer research from the 25 leading postcode areas with > 200 papers in 2009-11 and 2014-16, with distribution of major regional ethnicities. Contributions of other ethnic groups are not shown.

*Areas with foreign contribution (Forn) > 50% shown in **bold**, with < 30% in italics.*

<table>
<thead>
<tr>
<th>PC area</th>
<th>Post-town</th>
<th>UKI</th>
<th>EUR</th>
<th>SAS</th>
<th>ASI</th>
<th>ISL</th>
<th>EEU</th>
<th>Total</th>
<th>Forn, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>SW</td>
<td>London SW</td>
<td>681</td>
<td>220</td>
<td>123</td>
<td>87.6</td>
<td>49.6</td>
<td>42.8</td>
<td>1241</td>
<td>44.8</td>
</tr>
<tr>
<td>CB</td>
<td>Cambridge</td>
<td>688</td>
<td>209</td>
<td>60.8</td>
<td>99.5</td>
<td>24.2</td>
<td>47.6</td>
<td>1164</td>
<td>40.6</td>
</tr>
<tr>
<td>WC</td>
<td>London WC</td>
<td>598</td>
<td>294</td>
<td>93.4</td>
<td>55.4</td>
<td>41.0</td>
<td>43.8</td>
<td>1160</td>
<td>48.1</td>
</tr>
<tr>
<td>OX</td>
<td>Oxford</td>
<td>624</td>
<td>194</td>
<td>62.5</td>
<td>60.6</td>
<td>38.0</td>
<td>35.7</td>
<td>1045</td>
<td>39.8</td>
</tr>
<tr>
<td>M</td>
<td>Manchester</td>
<td>657</td>
<td>128</td>
<td>70.2</td>
<td>43.3</td>
<td>38.8</td>
<td>20.6</td>
<td>983</td>
<td>32.4</td>
</tr>
<tr>
<td>SE</td>
<td>London SE</td>
<td>374</td>
<td>164</td>
<td>82.7</td>
<td>48.3</td>
<td>49.9</td>
<td>31.7</td>
<td>776</td>
<td>51.3</td>
</tr>
<tr>
<td>B</td>
<td>Birmingham</td>
<td>425</td>
<td>82.2</td>
<td>75.3</td>
<td>33.9</td>
<td>32.8</td>
<td>20.5</td>
<td>690</td>
<td>38.0</td>
</tr>
<tr>
<td>SM</td>
<td>Sutton</td>
<td>415</td>
<td>145</td>
<td>34.6</td>
<td>35.4</td>
<td>13.4</td>
<td>25.6</td>
<td>686</td>
<td>39.2</td>
</tr>
<tr>
<td>W</td>
<td>London W</td>
<td>300</td>
<td>136</td>
<td>84.8</td>
<td>46.5</td>
<td>52.7</td>
<td>33.0</td>
<td>683</td>
<td>32.4</td>
</tr>
<tr>
<td>LS</td>
<td>Leeds</td>
<td>477</td>
<td>67.3</td>
<td>47.0</td>
<td>21.6</td>
<td>20.6</td>
<td>16.2</td>
<td>662</td>
<td>27.8</td>
</tr>
<tr>
<td>G</td>
<td>Glasgow</td>
<td>469</td>
<td>77.6</td>
<td>25.6</td>
<td>27.6</td>
<td>20.6</td>
<td>14.7</td>
<td>645</td>
<td>27.1</td>
</tr>
<tr>
<td>L</td>
<td>Liverpool</td>
<td>366</td>
<td>76.4</td>
<td>43.2</td>
<td>36.7</td>
<td>27.3</td>
<td>15.9</td>
<td>582</td>
<td>36.8</td>
</tr>
<tr>
<td>EC</td>
<td>London EC</td>
<td>326</td>
<td>104</td>
<td>45.2</td>
<td>29.2</td>
<td>23.5</td>
<td>35.8</td>
<td>582</td>
<td>43.3</td>
</tr>
<tr>
<td>EH</td>
<td>Edinburgh</td>
<td>417</td>
<td>68.7</td>
<td>14.6</td>
<td>20.7</td>
<td>10.8</td>
<td>24.0</td>
<td>565</td>
<td>26.0</td>
</tr>
<tr>
<td>BT</td>
<td>Belfast</td>
<td>432</td>
<td>15.6</td>
<td>9.0</td>
<td>20.3</td>
<td>9.3</td>
<td>5.8</td>
<td>513</td>
<td>12.1</td>
</tr>
<tr>
<td>CF</td>
<td>Cardiff</td>
<td>351</td>
<td>40.0</td>
<td>22.8</td>
<td>68.3</td>
<td>10.7</td>
<td>9.9</td>
<td>509</td>
<td>30.8</td>
</tr>
<tr>
<td>NG</td>
<td>Nottingham</td>
<td>315</td>
<td>36.2</td>
<td>62.4</td>
<td>26.1</td>
<td>35.5</td>
<td>20.0</td>
<td>507</td>
<td>37.8</td>
</tr>
<tr>
<td>S</td>
<td>Sheffield</td>
<td>331</td>
<td>41.4</td>
<td>30.5</td>
<td>15.7</td>
<td>14.6</td>
<td>7.5</td>
<td>451</td>
<td>26.4</td>
</tr>
<tr>
<td>NE</td>
<td>Newcastle</td>
<td>324</td>
<td>37.3</td>
<td>34.6</td>
<td>9.7</td>
<td>12.6</td>
<td>15.6</td>
<td>443</td>
<td>26.4</td>
</tr>
<tr>
<td>BS</td>
<td>Bristol</td>
<td>271</td>
<td>42.9</td>
<td>21.0</td>
<td>12.4</td>
<td>4.0</td>
<td>8.9</td>
<td>364</td>
<td>25.4</td>
</tr>
<tr>
<td>LE</td>
<td>Leicester</td>
<td>204</td>
<td>52.7</td>
<td>41.9</td>
<td>17.6</td>
<td>12.8</td>
<td>12.8</td>
<td>347</td>
<td>41.0</td>
</tr>
<tr>
<td>NW</td>
<td>London NW</td>
<td>166</td>
<td>63.0</td>
<td>41.2</td>
<td>23.4</td>
<td>15.9</td>
<td>12.5</td>
<td>331</td>
<td>49.6</td>
</tr>
<tr>
<td>SO</td>
<td>Southampton</td>
<td>222</td>
<td>50.4</td>
<td>17.5</td>
<td>10.6</td>
<td>12.8</td>
<td>6.7</td>
<td>325</td>
<td>31.6</td>
</tr>
<tr>
<td>DD</td>
<td>Dundee</td>
<td>180</td>
<td>37.0</td>
<td>17.4</td>
<td>13.3</td>
<td>6.5</td>
<td>9.2</td>
<td>272</td>
<td>32.8</td>
</tr>
<tr>
<td>E</td>
<td>London E</td>
<td>97.0</td>
<td>41.1</td>
<td>29.7</td>
<td>10.6</td>
<td>11.4</td>
<td>5.9</td>
<td>205</td>
<td>52.1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>11800</td>
<td>2834</td>
<td>1525</td>
<td>1021</td>
<td>748</td>
<td>624</td>
<td>19087</td>
<td>37.7</td>
</tr>
</tbody>
</table>

The ethnic minority contribution among these major research centres varied between 12% in BT (Northern Ireland) and 56% in London W (mostly Imperial College hospitals: Charing Cross, Hammersmith and St Mary's). It was also quite low (< 30%) in Scotland and the north of England. Figure 3 shows a map of the UK, shaded according to the percentage of ethnic minority cancer research contributions.
The South Asians were mainly in London (WC, W and SE) followed by Birmingham and Manchester; the West and East Europeans again in London (WC and SW) followed by Cambridge.
The sex distribution of UK cancer researchers

Table 5 shows the distribution by sex for authors from different world regions for both time periods together.

Table 5. The numbers and percentages of both males and females, the percentages that could be sexed, and the ratio of females to the sexable total, for authors of UK cancer research from each world region (for codes and countries represented, see Table 1).

<table>
<thead>
<tr>
<th>Region</th>
<th>Male</th>
<th>Female</th>
<th>blank</th>
<th>Total</th>
<th>M, %</th>
<th>F, %</th>
<th>Sexed, %</th>
<th>F/(M+F), %</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFR</td>
<td>470</td>
<td>345</td>
<td>237</td>
<td>1052</td>
<td>44.7</td>
<td>32.8</td>
<td>77.5</td>
<td>42</td>
</tr>
<tr>
<td>ASI</td>
<td>1522</td>
<td>1116</td>
<td>1744</td>
<td>4382</td>
<td>34.7</td>
<td>25.5</td>
<td>60.2</td>
<td>42</td>
</tr>
<tr>
<td>AUS</td>
<td>71</td>
<td>63</td>
<td>11</td>
<td>145</td>
<td>49.0</td>
<td>43.4</td>
<td>92.4</td>
<td>47</td>
</tr>
<tr>
<td>EEU</td>
<td>1211</td>
<td>1188</td>
<td>366</td>
<td>2765</td>
<td>43.8</td>
<td>43.0</td>
<td>86.8</td>
<td>50</td>
</tr>
<tr>
<td>EUR</td>
<td>6252</td>
<td>4732</td>
<td>1490</td>
<td>12474</td>
<td>50.1</td>
<td>37.9</td>
<td>88.1</td>
<td>43</td>
</tr>
<tr>
<td>ISL</td>
<td>1747</td>
<td>827</td>
<td>639</td>
<td>3213</td>
<td>54.4</td>
<td>25.7</td>
<td>80.1</td>
<td>32</td>
</tr>
<tr>
<td>LAT</td>
<td>233</td>
<td>264</td>
<td>71</td>
<td>568</td>
<td>41.0</td>
<td>46.5</td>
<td>87.5</td>
<td>53</td>
</tr>
<tr>
<td>NAM</td>
<td>38</td>
<td>52</td>
<td>10</td>
<td>100</td>
<td>38.0</td>
<td>52.0</td>
<td>90.0</td>
<td>58</td>
</tr>
<tr>
<td>SAS</td>
<td>3214</td>
<td>1619</td>
<td>1600</td>
<td>6433</td>
<td>50.0</td>
<td>25.2</td>
<td>75.1</td>
<td>33</td>
</tr>
<tr>
<td>UKI</td>
<td>23564</td>
<td>16800</td>
<td>6070</td>
<td>46434</td>
<td>50.7</td>
<td>36.2</td>
<td>86.9</td>
<td>42</td>
</tr>
<tr>
<td>WIN</td>
<td>40</td>
<td>19</td>
<td>15</td>
<td>74</td>
<td>54.1</td>
<td>25.7</td>
<td>79.7</td>
<td>32</td>
</tr>
<tr>
<td>(blank)</td>
<td>106</td>
<td>77</td>
<td>209</td>
<td>392</td>
<td>27.0</td>
<td>19.6</td>
<td>46.7</td>
<td>42</td>
</tr>
<tr>
<td>Total</td>
<td>38468</td>
<td>27102</td>
<td>12531</td>
<td>78101</td>
<td>49.3</td>
<td>34.7</td>
<td>84.0</td>
<td>41</td>
</tr>
</tbody>
</table>

The largest group was people with British or Irish names, and nearly 87% of them could be sexed. Of those that were, women contributed 42% of the cancer research papers. This ratio was lowest (32%) for people from the Islamic World, and almost as low for people from South Asia. It was highest for people from Latin America and the few with North American names, and only slightly less for those with Eastern European names.

In view of the efforts currently being made (Equality Challenge Unit, 2017; Munir et al., 2014) to attract more women into science and to facilitate their careers, we investigated whether there had been any change in the percentage of females / (males plus females), which was 41.3% overall (v.s.). This ratio had increased from 39.2% to 42.9% in five years, or by 9.4%. A more accurate representation is the change in the percentage of female contributions; this increased from 35.3% to 39.1%, or by nearly 11%. This is about a 2% growth per annum. The average contribution from both males and females went down from 2009-11 to 2014-16, but this may be an artefact because the same individual may be represented in several different formats and also work in more than one postcode area. For example, Abraham, Jean E. also signed papers as Abraham, Jean and as Abraham, J. E., and worked both in Cambridge (CB) and in London EC (probably Bart's Hospital) and had five separate entries in the list, and a total fractional contribution of 1.31 papers over the six years.

Discussion

That London contains the postcode areas with the highest proportions of members of ethnic minorities is not surprising as it has for a long time been a cosmopolitan city (The Economist, 2003). The 2011 Census (ONS, 2017) showed that it had 40% of ethnic minority people, of
whom just under half (18.5%) were Asian. In a report by the London Sustainable Development Commission (2012), London was found to perform well on the innovation quality of life indicator which may help to attract researchers from abroad.

There are a number of limitations to this study. The first concerns the ethnicity ascribed to the researchers by means of the OriginsInfo and other databases. There is inevitably a problem in the identification of people who change their surnames upon marriage. However, given names will not normally change and this is taken into account by the OriginsInfo software. There is also a problem with Latin American names, which are often indistinguishable from those of Portugal and Spain; these researchers have mostly been classified as western European. On the other hand, the UK contributions may also have been over-estimated because Australian, Canadian and US researchers working here will often have “British” or “Irish” names.

Although the fractional contributions of the different groups can be calculated, it is much harder to determine the numbers of individual researchers because of different name formats, and the possibility of some researchers, particularly senior people, having appointments in more than one postcode area. We have therefore not attempted to count the numbers of people but only the contributions of the main ethnic groups.

The OriginsInfo classification system is also unable to distinguish between recent immigrants, who are often described as “first generation” ethnic minorities, and people whose family has been in the UK for some time and who were born and educated here, and usually regard themselves as British, even if hyphenated. There is a possible way round this problem, if we assume that the division into first- and second-generation minorities among cancer researchers is similar to that among medical doctors of the same ethnicity, based on their names.

For a given ethnicity, e.g., Indian, we first counted the number of doctors who had qualified in India, and we then determined the ethnicity of all UK doctors who qualified in the UK by matching their surnames, first, to the list of ones classed as Indian by the OriginsInfo database, and second, to the list of surnames of UK doctors who qualified in India. These two numbers were then summed to give an estimated total for the number of ethnic Indians who qualified as doctors in the UK. The results for six countries are shown in Table 6.

Table 6. Estimates of the numbers and percentages of doctors of given ethnicity who qualified in their 'ethnic' country and in the UK. GMC = General Medical Council list. Our hypothesis is that this division is similar to that for cancer researchers of this ethnicity. For country codes, see Table 3.

<table>
<thead>
<tr>
<th>Country</th>
<th>Qualified in the UK</th>
<th>Abroad</th>
<th>Total</th>
<th>Abroad %</th>
<th>UK %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Origins Info names</td>
<td>Names from GMC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IN</td>
<td>7755</td>
<td>22259</td>
<td>32000</td>
<td>62014</td>
<td>51.6</td>
</tr>
<tr>
<td>LK</td>
<td>681</td>
<td>9862</td>
<td>3544</td>
<td>14087</td>
<td>25.2</td>
</tr>
<tr>
<td>PK + BD</td>
<td>4026</td>
<td>3711</td>
<td>11396</td>
<td>19133</td>
<td>59.6</td>
</tr>
<tr>
<td>SAS total</td>
<td>12462</td>
<td>35832</td>
<td>46940</td>
<td>95234</td>
<td>49.3</td>
</tr>
<tr>
<td>DE</td>
<td>5390</td>
<td>23250</td>
<td>6139</td>
<td>34779</td>
<td>17.7</td>
</tr>
<tr>
<td>PL</td>
<td>932</td>
<td>6706</td>
<td>3268</td>
<td>10906</td>
<td>30.0</td>
</tr>
<tr>
<td>IT</td>
<td>851</td>
<td>3505</td>
<td>4188</td>
<td>8544</td>
<td>49.0</td>
</tr>
</tbody>
</table>
It may seem strange that so many "Germans" were actually educated in the UK, but a check on the Web of 16 leading UK cancer researchers who had been characterised as German by OriginsInfo revealed that only five of them (31%) were actually Germans who had been educated there. [There are many UK citizens with German names whose forebears came to this country in the 19th and 20th centuries, many to escape persecution under the Nazis, and who went on to have distinguished careers in many walks of life.] A similar check on 11 leading "Italians" indicated that six of them were really Italians who had studied there. [There have historically been fewer Italian immigrants in the UK.]

We did similar checks on 37 leading researchers with South Asian names (25 "Indians", seven "Pakistanis", four "Sri Lankans" and one "Bangladeshi"). Of the 33 whose education could be found from their websites, 21 had received their doctorates in the UK and 12 overseas, so only 36% were first generation South Asian immigrants, compared with 49% estimated in Table 6. So it appears that the above table may over-estimate the numbers of first-generation researchers from South Asia, because many ethnic Indians have been settled here for several generations. [There was free entry for Commonwealth citizens in the early post-war decades, and many also arrived when they were expelled from Uganda in 1972.]

However, the table may under-estimate the numbers of first-generation Europeans who have come to the UK fairly recently, often as a result of European fellowship schemes such as the Marie Curie programme (European Commission, 2017), or from taking up jobs in cancer research advertised internationally.

**Conclusion**

The use of names, both surnames and given names, has allowed us to determine the contribution of ethnic minorities to UK cancer research, which is now about 40% and is rising slowly. The earlier paper (Begum et al., 2017) indicated that the research teams containing immigrants were publishing cancer papers in higher impact journals, but that there was little difference in their actual citations. At the time of writing, the effects of the UK’s departure from the European Union are not known, but the scientific community is concerned to ensure that researchers from other countries will still be welcome.

The use of postcodes for the addresses on the papers has shown that these ethnic minorities are concentrated in London and other major cities, but not in the rest of Great Britain or in Northern Ireland. Most of the researchers could be sexed from their given names, and this showed that women were still contributing less than half the total fractional count of UK cancer research papers (43% in 2014-16) but that this proportion was increasing by about 2% each year.

**Acknowledgements**

We are most grateful to the following who helped us by sexing the sets of given names from different world regions: for Africa, Victoria Bakaré of KCL, for Korea, Mijoo Hyun of Kew; for Eastern Europe, Argo Soon of the Estonian Research Council; for the Islamic World, Saleh Alessy of KCL and Zizette Said of Cairo; for South Asia, Aparna Basu of Delhi and Ajay Aggarwal of KCL; for Germany, Reinhold Wentz of Twickenham.
References


Multiple-keyword Co-occurrence Analysis and Time Series Prediction of Research Hotspots

Xiao Ming¹ and Xu Ye²

¹ming_xiao@bnu.edu.cn
Beijing Normal University, Beijing(China)

²xuye1219078794@qq.com
Beijing Normal University, Beijing(China)

Abstract
This study proposes that multiple-keyword co-occurrence can more accurately reveal the most frequently researched topics. The Apriori algorithm is used to mine the co-occurrence relation of keywords in order to predict the research hotspots by time series prediction. The keyword matrix of 15 core journals in library and information science in China is mined using related rules year by year via the Apriori algorithm. A total of 360 keyword sets are mined, including 1 five-keyword co-occurrence set, 6 four-keyword co-occurrence sets, and 39 three-keyword co-occurrence sets. Finally, 4 typical keyword sets are selected to predict their support, which will reflect the most widely researched topics for 2017.

Conference Topic
Co-occurrence analysis; Methods and techniques

Introduction
Callon, Courtial and Turner (1983) improved the theory and method of co-word analysis in the 1980s. Li and Ba (2017) considered that the statistical methods applied in co-term analysis dealing with word-related matrix mainly include co-term clustering analysis, co-term correlation analysis, co-term frequency analysis, and burst detection method. Li (2013) conducted a three-word co-occurrence analysis; regardless, studies have been rarely conducted on the coexistence between words from the perspective of multiple coexistence (Gao et al., 2014) and the multiple-keyword co-occurrence analysis derived from it. The current study presents the hypothesis that multiple-word co-occurrence analysis can reveal the research topic and content of a single thesis to more accurately reflect the research hotspot in this field.

The common methods revealing research hotspots by co-term analysis are multivariate statistical analysis and social network analysis (Tang and Zhang, 2014). In the multivariate statistical analysis, co-term clustering analysis and co-term correlation analysis are used to reveal the information of the correlation matrix. Apriori algorithm was first introduced into knowledge discovery based literatures by Hristovski et al (2000). Lei et al (2014) applied the association rules mining in
co-term analysis field and revealed the relationship between double phrases and ternary phrases in the target domains effectively. However, the research of multiple-word co-occurrence is still in its infant stage, which naturally confronts with lots of challenges, problems, and barriers during its maturation. The Apriori algorithm is used in the current study to mine the co-occurrence relation of the keywords. The most frequently researched topics in Library and Information Science are predicted in the time series.

Data and Methodology

Data Sources
To predict the research hotspots in library and information science, keywords from 15 Chinese journals were collected over a 10-year period from the China National Knowledge Infrastructure (CNKI) full-text database. These journals included the Journal of Library Science in China, Journal of Academic Libraries, etc. A total of 172,002 keywords were retrieved from 44,648 papers.

Data preprocessing
Co-term analysis uses accurate data as its basis. The collection of keywords obtained from the CNKI database includes many inconsistencies, which are not conducive to computer reading. Thus, data cleaning was necessary. In this study, data preprocessing employed manual processing, which mainly consists of the following steps: (1) deletion of unrelated data; (2) merging of synonymous Chinese and English keywords; (3) merging of singular and plural cases of English keywords; (4) merging of synonymous Chinese keywords; (5) separation of each keyword by a uniform delimiter.

Methodology
1. Multiple-keyword co-occurrence
Keywords are a collection of words that can be used to express the main knowledge structure and core knowledge content of a single study. The co-occurrence between two keywords can be referred to as a two-keyword co-occurrence, and that between three keywords is referred to as three-keyword co-occurrence, and so on. When a number of keywords appear simultaneously in a study, and the frequency of occurrence in the sample is very high, the field of study represented by these keywords is considered a widely researched area for a certain period. The higher the number of keyword co-occurrence, the more accurate the description of the research area represented.

2. Apriori algorithm
The Apriori algorithm is the most commonly used algorithm for mining frequent item sets with association rules. The core idea is to generate candidates and their support through a connection and then generate frequent item sets by pruning. The algorithm, which is currently applied in co-term correlation analysis, comprises four basic concepts: support, confidence, minimum support, and minimum confidence. Support is the probability that two keywords A and B appear simultaneously in a
where $Total\_count(A)$ refers to the number of all keywords sets, and $Support\_count(A \cup B)$ refers to the number of keyword sets that keyword $A$ and $B$ appeared simultaneously. Confidence is the probability that keyword $A$ appears at the same time keyword $B$ appears:

$$Confidence(A \Rightarrow B) = P(A | B) = \frac{Support\_count(A \cup B)}{Support\_count(A)}$$

where $Support\_count(A)$ refers to the number of keyword sets that have keyword $A$.

The minimum support is a threshold measuring the degree of support defined by the expert, indicating the minimum importance of the keyword set in the statistical sense. The minimum confidence is a threshold measuring confidence defined by experts, representing the lowest reliability of the association rule. Rules that satisfy the minimum support threshold and the minimum confidence threshold are called strong rules.

3. Time series prediction

Time series prediction compiles and analyzes time series according to the time series reflected in the development process, direction and trends. An analogy or extension is then made to predict the level that may be reached in the subsequent or later periods. This study employs polynomial fitting prediction, which aims to find a set of polynomial coefficients $a_i, i=1,2,\cdots,n+1$ that provides a better fit between the polynomial $\psi(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ and the original data and results in a smaller overall fitting error.

Results and Discussion

Multiple-keyword co-occurrence analysis

1. Setting the threshold of minimum support and minimum confidence

The setting of threshold about the minimum support and the minimum confidence can directly affect the number and reliability of the mined keyword set. Hu and Chen (2014) estimated the covariance ratio of the keywords in the field to be 0.73%. Given that semantic association and background knowledge between the words are neglected, the minimum support is set to 0.001, and the minimum confidence is set to 0.5.

2. Keyword set analysis

The Apriori algorithm is used to mine the association rules for keyword matrix each year, and a total of 360 keyword sets are mined. The sets include a five-keyword co-occurrence set, 6 four-keyword co-occurrence sets, and 39 three-keyword co-occurrence sets. The specific scenario is presented in Tables 1 and 2:

**Table 1. Keyword sets sorted by number of words (top 10).**
As shown in Table 1, outsourcing, MARC, information organization, metadata, and library catalogue are often bundled in keyword sets. Information organization, MARC, and metadata technology are applied in the library catalogue, and library catalogue develops outsourcing. Museum science, library science, and archive science are often bundled, which leads us to speculate that with the development of associated data, the concept and implementation of integration are moving forward. Big data, information engineering, and intelligent city are coming together, indicating that the construction of an intelligent city will be another important manifestation.

Table 2 shows that the college library was a research hotspot from 2007 to 2009. Building a knowledge map by CiteSpace was a research hotspot from 2012 to 2016. The support for keyword sets is generally low. While co-term association analysis

<table>
<thead>
<tr>
<th>ID</th>
<th>Year</th>
<th>Keyword Set</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2015</td>
<td>Outsourcing–MARC–Information organization–Metadata–Library catalogue</td>
<td>0.00104</td>
<td>0.80</td>
</tr>
<tr>
<td>2</td>
<td>2013</td>
<td>Librarian–Empirical research–Future career–Stereotype</td>
<td>0.00120</td>
<td>0.83</td>
</tr>
<tr>
<td>3</td>
<td>2015</td>
<td>Outsourcing–MARC–Information organization–Metadata</td>
<td>0.00104</td>
<td>0.80</td>
</tr>
<tr>
<td>4</td>
<td>2015</td>
<td>Outsourcing–MARC–Information organization–Library catalogue</td>
<td>0.00104</td>
<td>0.80</td>
</tr>
<tr>
<td>5</td>
<td>2015</td>
<td>Outsourcing–MARC–Metadata–Library catalogue</td>
<td>0.00104</td>
<td>0.80</td>
</tr>
<tr>
<td>6</td>
<td>2015</td>
<td>Library catalogue–MARC–Information organization–Metadata</td>
<td>0.00104</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>2015</td>
<td>Outsourcing–Information organization–Metadata–Library catalogue</td>
<td>0.00104</td>
<td>0.80</td>
</tr>
<tr>
<td>8</td>
<td>2015</td>
<td>Subject librarian–University library–Subject service</td>
<td>0.00130</td>
<td>0.83</td>
</tr>
<tr>
<td>9</td>
<td>2016</td>
<td>Museum–Library–Archives</td>
<td>0.00114</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>2012</td>
<td>Big data–Information engineering–Intelligent city</td>
<td>0.00114</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Table 2. Top 10 keyword sets sorted by support.
ignores the semantic relation between keywords, it also reflects that the research
direction of Chinese Library and Information Science field has been fragmented in the
recent 10 years and that research results develop in different areas.

**Time series prediction of research hotspots in library and information science in China**

At present, the co-term analysis applied to recognize research hotspots covers long
periods. Few people predict possible hotspots by using historical data. In this study,
research hotspots are analyzed year by year, and potential research hotspots are
explored by time series prediction.

Table 3 presents the highest support received by the keyword sets, arranged by year.

<table>
<thead>
<tr>
<th>ID</th>
<th>Year</th>
<th>Keyword Set</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2007</td>
<td>Library–College</td>
<td>0.00307</td>
<td>0.78</td>
</tr>
<tr>
<td>2</td>
<td>2008</td>
<td>Library–Core value</td>
<td>0.00339</td>
<td>0.76</td>
</tr>
<tr>
<td>3</td>
<td>2009</td>
<td>Library–College</td>
<td>0.00308</td>
<td>0.80</td>
</tr>
<tr>
<td>4</td>
<td>2010</td>
<td>Library–College</td>
<td>0.00207</td>
<td>0.61</td>
</tr>
<tr>
<td>5</td>
<td>2011</td>
<td>Public library–Free open</td>
<td>0.00298</td>
<td>0.79</td>
</tr>
<tr>
<td>6</td>
<td>2012</td>
<td>CiteSpace–Knowledge mapping</td>
<td>0.00337</td>
<td>0.58</td>
</tr>
<tr>
<td>7</td>
<td>2013</td>
<td>Mobile library–Mobile service</td>
<td>0.00311</td>
<td>0.52</td>
</tr>
<tr>
<td>8</td>
<td>2014</td>
<td>Library–Copyright</td>
<td>0.00261</td>
<td>0.56</td>
</tr>
<tr>
<td>9</td>
<td>2015</td>
<td>Document delivery–Interlibrary loan</td>
<td>0.00234</td>
<td>0.69</td>
</tr>
<tr>
<td>10</td>
<td>2016</td>
<td>Subject librarian–Subject service</td>
<td>0.00337</td>
<td>0.68</td>
</tr>
</tbody>
</table>

According to the specific scenario of research hotspots for each year, some
representative keyword sets are selected for analysis and prediction. These sets
include keyword set1 (CiteSpace–knowledge mapping), keyword set2
(library–college), keyword set3 (cloud services–cloud computing), and keyword set4
(museum–library–archives). The support corresponding to each set is plotted in
Figure 1.

![Figure 1. Age distribution of the support of some keyword sets.](image)

The keyword sets are grouped into 4 time series by age distribution. Polynomial
fitting is performed to make a prediction. The specific scenarios are depicted in Figure
The keywords sets and their corresponding support for 2017 are as follows: keyword set1 (CiteSpace–knowledge mapping), 0.2787%; keyword set2 (library–college), 0%; keyword set3 (cloud services–cloud computing), 0.2317%; and keyword set4 (museum–library–archives), 0.0265%. Cloud services–cloud computing has the potential to be a research hotspot, which mainly discusses how to build a cloud service model in the cloud computing environment. CiteSpace–knowledge mapping is supposed to continue to be a hot subject of research in the area of scientometrics. Visualizing and analyzing trends and patterns in scientific literatures by using CiteSpace software will still to be a research hotspot in the field of library and information science in China. The research results indicate that library–college will not to be a research hotspot. The research on information service and resources construction of college library has been perfected. A small number of researchers are expected to remain interested in researching the digital libraries and archives.

**Conclusions**

This study proposes multiple-keyword co-occurrence analysis, which can show the research topic of the study to provide a more accurate result for research hotspots. By using the Apriori algorithm, the keyword matrix of 15 core journals in library and information science in China is mined with related rules year by year; 360 keyword sets are mined. Polynomial fitting is then used to predict the support corresponding to each of the four typical keyword sets for a certain period; the most popular research topic for 2017 is predicted.

This study has successfully mined keyword sets for co-term analysis and provides a short-term forecast by time series forecasting. Regardless, data disambiguation and short-term forecasting can still be improved. In addition, the semantic correlation between keywords should be ignored. This study sets the future research direction toward the combination of the semantic correlation between the keywords themselves to construct richer association rules.
Acknowledgments
This work was supported by a grant from the National Social Science Foundation of China (No. 16BTQ073).

References
European newspaper reports of non-communicable disease research, 2002-13

Elena Pallari¹ and Grant Lewison²

¹ elena.pallari@kcl.ac.uk
King’s College London, Institute of Psychiatry, Psychology & Neuroscience, De Crespigny Park, Denmark Hill, London, SE5 8AF, (UK)

² grantlewison@aol.co.uk
King’s College London, Cancer Studies, Guy’s Hospital, Great Maze Pond, London SE1 9RT (UK)

Abstract
This project was part of a large mapping exercise on the research on five non-communicable diseases in Europe in 2002-13. This involved the identification and analysis of over 700,000 papers and also examination of their impact through citations in other papers, on clinical guidelines and in European papers. Stories in newspapers inform many groups, from politicians to the general public, about progress in medical research, but they can also report selectively and so bias our understanding of the state of play. We employed a large multi-lingual team to search the archives of 32 newspapers in 21 countries, and compiled an archive of over 9000 stories and the papers that they cited. These were mostly about cancer, how to prevent it rather than treat it, and over-cited European research but neglected that from east Asia. Some papers were cited in many different newspapers and so achieved wide coverage. Commentators’ views on the importance of the research were sought on 13% of the continental newspaper stories, but 37% of those in the UK Daily Mail. The latter paper mostly quoted UK charities who thereby gained free publicity.

Conference Topics
Indicators; science policy and research assessment; science communication; the application of informetrics on evaluation

Introduction
There is extensive literature about the stories of medical research in newspapers, but almost all of them concern only a single country and one type of research (e.g., cancer or genetics). Studies have been carried out in the UK (Entwistle, 1995; Robinson et al., 2013); in the USA (Burns et al., 1995; Stryker, 2002); Russia (Lichterman, 2002); the Netherlands (Hijmans, Pleijter & Wester, 2003); on prostate cancer screening in Australia (MacKenzie et al., 2007); on healthy lifestyles and hypertension in Canada (Hayes et al., 2007; Collin & Hughes, 2011); on diabetes in Japan (Akamatsu, Naito & Nakayama, 2009); on male circumcision in Kenya (Muzyka et al., 2012); in Brazil (Teixeira et al., 2012); and on medical uses of cannabis in Israel (Lewis, Broitman & Sznitman, 2015). An exception is the study of the newspaper coverage of the Severe Acute Respiratory Syndrome (SARS) in 2003-04 (Lewison, 2008) where news of the research that was hurriedly started was analysed in 15 news media from seven countries. The present paper is the first to examine the variation in coverage between 21 different European countries, with a combined total of 15 languages, and five distinct non-communicable disease (NCD) areas. These are cardiovascular disease including stroke, diabetes, mental disorders, cancer and respiratory diseases. Each of these was further subdivided into different disease areas, and the types of research were also analysed. So the study affords a wide view of how the European public is informed about medical research.
Despite falling circulations in many countries, newspapers are still very influential as news sources. Large headlines about medical research on the front page can grab the attention of many types of reader – politicians, their senior advisors, healthcare administrators, medical personnel, other researchers and the general public. They have the advantage over the social media of relative permanence so that analyses are reproducible. Moreover, many newspapers maintain searchable archives, or they can be searched through full-text databases such as Factiva or ProQuest NewsStand.

Newspaper stories have some of the characteristics of journal articles, in that they have a source (country and newspaper, date), an author, a title (headline), one or more citations, and (sometimes) commentators who put the new research in context. Reports of research in the mass media can also lead to higher citation scores for the papers cited in the stories (Phillips et al., 1991; Lewison et al., 2008; Fanelli, 2013).

Our objectives were to examine the balance of news coverage between the five NCDs in the different countries, and where appropriate, the balance of individual disease areas and also the types of research being reported. We also sought to investigate the extent of countries’ over-citation of their own researchers. This also occurs in mainstream research outputs (Lewison & Roe, 2012), and in the references on clinical guidelines (Grant et al., 2000), but we wondered whether journalists would be even more chauvinistic than scientists, and over-cite their own countrymen to a greater extent. Finally, we wanted to see the extent to which external commentators were brought in by the journalists to put the recent research results in context.

**Methodology**

*Selection of newspapers.*

Our original intention was to cover about 60 newspapers in all the 31 European countries (the 28 European Union Member States (MS) plus Iceland, Norway and Switzerland). There would be three newspapers in the large MS such as Germany and Italy, two in each of the medium-sized countries such as Austria, Belgium and Portugal, and one in the small countries. The selection was intended to cover both right- and left-wing newspapers, and “broadsheets” and “tabloids”, though these terms are now no longer an accurate reflection of the market being served. However the work proved to be more time-consuming than we had expected, so we perforce needed to restrict our coverage to some 31 newspapers in 22 countries, see Table 1. The trigraph (three-letter) codes for the individual papers are composed of the countries’ ISO codes and a letter to indicate the paper title.

*Selection of stories.*

In order to collect the relevant stories in these different newspapers, we assembled a team of some 31 recorders, with the appropriate language skills (see the Acknowledgement). Some were from our five European partners in the EU project; some were recruited from among King’s College London graduate students; and there were three volunteers from Cyprus, Hungary and Sweden who participated in this project and in return were given detailed data on NCD research in their countries.

All the recorders were brought to KCL for briefing and training by EP. They were shown how to select the stories, either from the newspaper archives or from the Factiva database © Dow Jones, by means of five search statements such as the one for cancer:

\[
(cancer \text{ OR } \text{leukaemi}^* \text{ OR } \text{melanoma}^* \text{ OR } \text{lymphoma}^*) \text{ AND (research}^* \text{ OR study OR scientists OR expert}^*)
\]
This, of course, had to be translated into the languages of the individual newspapers. The recorders then had to read the stories to check that they were citing research papers in journals—many were not, or the details given were inadequate to identify the paper, and were discarded. All relevant details were copied and pasted to an Excel spreadsheet, and all recorders used exactly the same format, even if some of the data were not present. The data included the source (codes as in Table 1), the date, the name(s) of the journalists, and codes to characterise the NCD (CARDI, DIABE, MENTH, ONCOL or RESPI), the disease area (e.g., cancer site or mental disorder), and the type of research (e.g., drug treatment, epidemiology). The recorders also tabulated the names of the cited scientists, their institutions and the journal in which the research was published, and the details of any commentators who were cited.

**Identification and analysis of cited papers.**

The recorders were then instructed to try and identify the research papers that had been cited in the stories in the Web of Science Core Collection (WoS), and to download their full bibliographic details to a series of text files. All document types could be included though in practice most were "articles". The files were converted to Excel spreadsheets by means of a macro (written by Philip Roe of Evaluametrics Ltd) and then the details of individual papers were copied across to the spreadsheet of stories. Only stories with papers in the WoS were retained for analysis.

The individual spreadsheets were then all combined to create a master sheet with over 8800 stories and their cited papers. These papers were then analysed by their addresses, with fractional country counts, and for cancer and mental health papers, by their type of research, although this was also coded by the recorders from the stories. Our main interest was in the European research papers, and the contributions of European countries were compared with their presence in NCD research in the years 2002-13, obtained from the mapping exercise.

**Table 1.** List of countries and newspapers whose NCD stories were included in the archive.

<table>
<thead>
<tr>
<th>Country</th>
<th>Code</th>
<th>Newspaper</th>
<th>Country</th>
<th>Code</th>
<th>Newspaper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>AT P</td>
<td>Die Presse</td>
<td>Italy</td>
<td>IT C</td>
<td>Corriere della Sera</td>
</tr>
<tr>
<td>Belgium</td>
<td>BE D</td>
<td>De Standaard</td>
<td>Italy</td>
<td>IT R</td>
<td>La Repubblica</td>
</tr>
<tr>
<td>Belgium</td>
<td>BE S</td>
<td>Le Soir</td>
<td>Netherlands</td>
<td>NL D</td>
<td>Het Algemeen Dagblad</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>BG D</td>
<td>Дневник (Dnevnik)</td>
<td>Netherlands</td>
<td>NL T</td>
<td>De Telegraaf</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>BG T</td>
<td>Труд (Trud)</td>
<td>Poland</td>
<td>PL F</td>
<td>Fakt</td>
</tr>
<tr>
<td>Croatia</td>
<td>HR V</td>
<td>Vecernji List</td>
<td>Portugal</td>
<td>PT C</td>
<td>Correio da Manhã</td>
</tr>
<tr>
<td>Cyprus</td>
<td>CY M</td>
<td>Cyprus Mail</td>
<td>Portugal</td>
<td>PT J</td>
<td>Jornal de Noticias</td>
</tr>
<tr>
<td>Czech Rep.</td>
<td>CZ B</td>
<td>Blesk</td>
<td>Romania</td>
<td>RO A</td>
<td>Adevarul</td>
</tr>
<tr>
<td>Denmark</td>
<td>DK J</td>
<td>Jyllands Posten</td>
<td>Spain</td>
<td>ES A</td>
<td>ABC</td>
</tr>
<tr>
<td>Estonia</td>
<td>EE O</td>
<td>Õhtuleht</td>
<td>Spain</td>
<td>ES M</td>
<td>El Mundo</td>
</tr>
<tr>
<td>Estonia</td>
<td>EE P</td>
<td>Postimees</td>
<td>Spain</td>
<td>ES P</td>
<td>El País</td>
</tr>
<tr>
<td>Finland</td>
<td>FI H</td>
<td>Helsingin Sanomat</td>
<td>Sweden</td>
<td>SE D</td>
<td>Svenska Dagbladet</td>
</tr>
<tr>
<td>France</td>
<td>FR M</td>
<td>Le Monde</td>
<td>Switzerland</td>
<td>CH B</td>
<td>Berner Zeitung</td>
</tr>
<tr>
<td>Germany</td>
<td>DE S</td>
<td>Süddeutsche Zeitung</td>
<td>UK</td>
<td>UK D</td>
<td>Daily Mail</td>
</tr>
<tr>
<td>Greece</td>
<td>GR V</td>
<td>To Bijua (To Bema)</td>
<td>UK</td>
<td>UK G</td>
<td>The Guardian</td>
</tr>
<tr>
<td>Hungary</td>
<td>HU M</td>
<td>Magyar Nemzet</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: the first part of the codes is the ISO2 country code.*
Results

Comparison of NCD news coverage with disease burden.

Figure 1 shows the distribution of news stories by NCD for the complete set of papers and for the nine countries with at least 300 stories: Belgium, Germany, Italy, Netherlands, Portugal, Romania, Spain, Sweden and the UK. [These are not necessarily the countries with the greatest overall coverage of medical research in the popular press.] The percentages mostly sum to more than 100% because some stories described research affecting more than one NCD, for example rising obesity will increase both diabetes and heart disease, and smoking affects both cancer and Chronic Obstructive Pulmonary Disease (COPD), a respiratory disease.

Is this a fair representation of the relative burden from the five diseases? This is conveniently measured in Disability-Adjusted Life Years, which takes account not only of early death (compared with life expectancy in Japan), but also of disability. Figure 2 shows the pattern for the five NCDs as a group, based on data published by the World Health Organization for 2012 (WHO, 2014).

One would hardly expect these two charts to be in synchrony, but there are some notable differences between them. First, cancer (ONCOL) is the most reported story in all countries except for Romania (RO) and Sweden (SE), where cardiovascular research (CARDI) gets more coverage. This is as expected in Romania where it accounts for more than half of the total disease burden from the five NCDs as a group, and it also causes a higher burden than cancer in Sweden – as it does in Europe as a whole – but overall there are over 70% more ONCOL stories than CARDI ones.

![Figure 1. Distribution of newspaper stories among the five NCDs (cardiovascular disease, diabetes, mental disorders, oncology and respiratory diseases) for nine individual countries (ranked by total number of stories) and all 22 countries in Europe (see Table 1).](image)

Diabetes research (DIABE) gets much more coverage in newspapers (11%) than its current burden (4.2% of all DALYs for the five NCDs, and 2.6% of DALYs for all diseases) would suggest, but the burden is increasing rapidly, largely because of increasing obesity, so this may be a sensible amount of coverage in view of the need for citizens to take more care of
their health. On the other hand, respiratory disease research (RESPI) is under-reported, with 6.5% of stories compared with 8.2% of DALYs, and particularly in the UK (6.8% of stories, 12.7% of DALYs) and Portugal (3.0% of stories, 9.2% of DALYs).

Figure 2. Distribution of disease burden from five NCDs in 2012 among nine European countries and in all 31, DALYs, all ages, both sexes.

Mental disorders research (MENTH) is mostly reported proportionately to its relative burden, except in Sweden and Spain, which feature few MENTH stories. But Romania has many such stories, despite having a low burden (13% of all DALYs compared with the European average of 23%).

Reportage of individual diseases within each NCD.
The five NCDs each cover a number of different diseases, and the next analysis shows how they compare with the European burden from each. In cardiovascular disease research (CARDI), the two main areas are ischaemic heart disease and cerebrovascular disease, or stroke. In 2012, these accounted for 14% and 8% of the DALYs for the five NCDs, but the numbers of stories were only 11% and 4%, so stroke research, in particular, was somewhat neglected by the journalists.

We selected five mental disorders for analysis: addiction (ADD), alcoholism (ALC), Alzheimer’s disease and other dementias (ALZ), anxiety (ANX) and unipolar depression (DEP). Figure 3 shows the percentages of DALYs, of European research, and of newspaper stories, all relative to the total for the five NCDs. Figure 4 gives data for eight leading cancer sites.
Figure 3. Five mental disorders in Europe: disease burden (DALYs), research and newspaper stories; all as % of total for five NCDs.

Figure 4. Eight cancer sites in Europe: disease burden (DALYs), research and newspaper stories; all as % of total for five NCDs.

Figure 3 shows that the amount of research in these five mental disorders is not large compared with the disease burden, particularly for alcoholism and depression. On the other hand, the amount of newspaper coverage is good for Alzheimer's disease and other dementias, where the burden is increasing because of greater longevity. Alcoholism is a serious problem in Europe (Rajendram et al., 2006), and is ranked as one of the three main risk factors in the Healthgrove website for almost half of the 188 countries listed, and for 30 of the 31 European countries (all except Malta).

For the different manifestations of cancer, Figure 4 shows that lung cancer is under-researched in relation to its burden, together with pancreatic cancer whose incidence is increasing. The news coverage is reasonable for most cancers, but seems excessive for breast cancer, melanomas and prostate cancer. These seem to be the ones of most concern to people, both women and men, so it is perhaps not surprising.
Coverage of research on the last NCD, respiratory diseases (RESPI), has already been presented (Pallari et al., 2016). They found that asthma, although much less burdensome than COPD, was nevertheless researched much more and reported even more, probably because the sufferings of asthmatic children are more newsworthy than those of old smokers.

The types of research that were reported.

We employed four-character codes to designate research types; they were not the same for the different NCDs but most were common to them all. Overall, the research types were as shown in Table 2. Most of the stories involving surgery were on cancer, as were all of those on radiotherapy. The sum of the percentages is greater than 100% because many of the stories covered two (or more) research types, but some could not be coded in this way.

Table 2. Types of research that were mainly cited in European newspaper stories about NCD research, 2002-13.

<table>
<thead>
<tr>
<th>Type of research</th>
<th>Code</th>
<th>Stories</th>
<th>%</th>
<th>Type of research</th>
<th>Code</th>
<th>Stories</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epidemiology</td>
<td>EPID</td>
<td>1831</td>
<td>20.7</td>
<td>Other treatments</td>
<td>OTHT</td>
<td>339</td>
<td>3.8</td>
</tr>
<tr>
<td>Genetics</td>
<td>GENE</td>
<td>1461</td>
<td>16.5</td>
<td>Prognosis</td>
<td>PROG</td>
<td>329</td>
<td>3.7</td>
</tr>
<tr>
<td>Lifestyle choices</td>
<td>LIFE</td>
<td>1327</td>
<td>15.0</td>
<td>Quality of life</td>
<td>QUAL</td>
<td>295</td>
<td>3.3</td>
</tr>
<tr>
<td>Drugs/chemotherapy</td>
<td>DRUG</td>
<td>1312</td>
<td>14.9</td>
<td>Screening</td>
<td>SCRE</td>
<td>234</td>
<td>2.6</td>
</tr>
<tr>
<td>Nutrition/foods</td>
<td>NUTR</td>
<td>1223</td>
<td>13.8</td>
<td>Toxicology</td>
<td>TOXI</td>
<td>217</td>
<td>2.5</td>
</tr>
<tr>
<td>Pathology</td>
<td>PATH</td>
<td>702</td>
<td>7.9</td>
<td>Surgery</td>
<td>SURG</td>
<td>129</td>
<td>1.5</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>DIAG</td>
<td>480</td>
<td>5.4</td>
<td>Radiotherapy</td>
<td>RADI</td>
<td>101</td>
<td>1.1</td>
</tr>
</tbody>
</table>

It is striking that so many of the stories reported research on how the NCDs could be avoided (EPID, LIFE and NUTR) and that treatments (most involving drugs) were very much a secondary concern. For cancer, the main treatments described were of new drugs used in chemotherapy, some used for immunotherapy or targeted therapy; surgery and radiotherapy (which actually can cure cancer) were rarely mentioned. This is a serious matter as the public are thereby persuaded that the provision of expensive drugs is the way to improve cancer survival, and the case for better radiotherapy equipment is not appreciated by either the public or politicians.

The countries whose research was cited by the newspaper stories.

The addresses of the papers cited in the newspaper stories were analysed on a fractional count basis for the countries. It was immediately apparent that European countries were over-represented in these papers compared with their presence in research in the five NCDs for the years 2002-13: the over-citation ratios varied between x 1.19 in RESPI and x 1.51 in ONCOL. But some countries' research was much better cited than others’, as Table 3 shows.

Research from the five Nordic countries is clearly very well-cited in all five NCDs with minor exceptions (notably Norwegian work in diabetes). It is particularly striking that Icelandic research attracts so much attention, especially in cancer and mental health. Even though the numbers of papers are small (6.4 and 5.1, respectively) they are statistically significant (on the Poisson distribution) with p << 0.001%. On the other hand, research from central European countries (Austria, Germany, Hungary, Italy, Poland and even Switzerland) is under-cited compared with the countries' presence in these NCDs. The same is true for research from east Asian countries (China, Japan, Korea and Taiwan) even though most of it is published in English.

A possible explanatory factor is that the coverage of NCD research in our archive of stories is very unequal between countries, and one would expect that newspapers would preferentially
cite research from their own researchers. This is the case, and small countries (in terms of scientific output) over-cite their own research by a bigger factor than do large countries, see Figure 5; the comparison is with the countries' presence in the five NCDs as a group.

Table 3. Ratios of observed to expected country presence on research papers cited by European newspaper stories, 2002-13, in five NCDs. Note: codes have been shortened to save space.

Values > 2.0 bold 14 pt; > 1.41 bold 12 pt; < 0.71 italic 12 pt; < 0.5 italic 10 pt.

| ISO2 | CAR | DIA | MEN | ONC | RES | ISO2 | CAR | DIA | MEN | ONC | RES |
|------|-----|-----|-----|-----|-----|------|-----|-----|-----|-----|-----|-----|
| UK   | 3.24| 2.65| 2.25| 3.56| 1.80| NO   | 1.47| 0.69| 1.62| 2.55| 1.28|
| NL   | 1.56| 2.05| 1.71| 2.18| 1.69| CH   | 0.92| 0.67| 0.97| 0.83| 1.15|
| SE   | 3.06| 1.96| 2.21| 2.90| 1.15| PT   | 0.41| 0.92| 2.16| 3.70| 0.71|
| FR   | 0.77| 1.02| 0.92| 1.12| 0.31| GR   | 0.77| 0.46| 0.57| 0.50| 0.94|
| DE   | 0.55| 0.39| 0.51| 0.45| 0.79| AT   | 0.28| 0.44| 0.23| 0.53| 1.31|
| IT   | 0.68| 0.63| 0.52| 0.74| 0.55| IE   | 1.10| 0.31| 0.45| 0.92| 0.54|
| ES   | 0.91| 0.99| 0.69| 1.47| 1.03| PL   | 0.12| 0.02| 0.17| 0.38| 0.11|
| FI   | 2.97| 2.69| 3.55| 2.67| 2.33| IS   | 3.79| 3.62| 6.47| 7.56| 0.85|
| DK   | 3.09| 1.02| 1.74| 2.76| 3.86| HU   | 0.46| 0.34| 0.50| 1.00| 0.23|
| BE   | 1.41| 2.26| 1.06| 2.71| 1.45| RO   | 4.04| 1.21| 0.49| 0.19| 0.00|

Figure 5. Over-citation ratio of own country papers by European newspaper stories about NCD research (countries with 160 or more stories only), fractional country counts.

Figure 5 shows that the Nordic countries over-cite their own countrymen more than expected, and this may be one reason why their papers are so well cited (Table 3). Conversely, German journalists seem relatively reluctant to cite German research, and this may be a factor in the low newspaper citation scores of German papers in Table 3, particularly in comparison with the UK, whose presence in research in the five NCDs as a group is similar to that of Germany.
*Highly-cited papers*

In conventional citation analysis, papers are often ranked by the number of times that they have been cited by other research papers. This should be in a fixed time window, often three or five years. There are also papers that receive many cites in newspapers, and if we had been able to process all European daily newspapers, it is likely that the citation scores of NCD papers would have been much greater. It is still worthwhile to examine the distribution of citation numbers, but it is not really necessary to use a particular time window because most stories report on new research as soon as it is published, or even beforehand if the paper appears online prior to the print edition of the journal. One paper received 18 newspaper cites and two others, 10; two were in cancer and one in cardiovascular research. The distribution for the other papers cited twice or more in our selection of newspapers is shown in Figure 6. It is well fitted by means of an exponential relationship.

![Figure 6. Distribution of numbers of citations to NCD research papers in European newspaper stories, 2002-13.](image)

*Commentators*

Some of the newspaper stories attempted to put the research findings into context by seeking the views of a knowledgeable commentator. For the continental European newspaper stories, only 741 did so (12.6%), but for the UK *Daily Mail*, 739 of its stories had a commentator (37%) and for *The Guardian* 204 had one (25%). The commenting organisations were coded with their sector (government, private-non-profit, industry; the PNP sector was further divided into charities, foundations, hospitals, universities and other non-profits) and their country. Table 4 shows the main parameters of these organisations for the stories from 21 continental countries (many of which appeared to have no commentators) and for the two UK papers, *Daily Mail* and *The Guardian*.

The big difference between the two is that the *Daily Mail* and *The Guardian* sought the opinion of charities far more often than did the continental journalists. Most of these were UK-based, such as Cancer Research UK (186 mentions), the British Heart Foundation (99), and the Alzheimer's Society (70). (UK charities were also quoted 42 times in the continental newspaper stories.) The two UK papers also sought the opinion of over 75 other non-profit organisations, including several Royal Colleges of professional medical personnel and other professional associations. There were also some lobbying groups such as Friends of the Earth, JABS (Justice Awareness and Basic Support – for victims of vaccination), and the
Salmon Farm Protest Group. For the continental papers, the main source of comments was universities, especially in Sweden (61 mentions), Belgium (52) and the USA (44).

Table 4. Types of commenting organisations on NCD research stories in the newspapers of 21 continental European countries (Continental) and in the UK Daily Mail.

<table>
<thead>
<tr>
<th>Organisation</th>
<th>Continental newspapers</th>
<th>Daily Mail &amp; The Guardian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Organisations</td>
<td>Percent</td>
</tr>
<tr>
<td>Total</td>
<td>722</td>
<td>12.3</td>
</tr>
<tr>
<td>Government</td>
<td>121</td>
<td>2.1</td>
</tr>
<tr>
<td>Private-non-profit</td>
<td>539</td>
<td>9.2</td>
</tr>
<tr>
<td>Charities</td>
<td>78</td>
<td>1.3</td>
</tr>
<tr>
<td>Foundations</td>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>Hospitals</td>
<td>128</td>
<td>2.2</td>
</tr>
<tr>
<td>Universities</td>
<td>221</td>
<td>3.8</td>
</tr>
<tr>
<td>Other NP</td>
<td>109</td>
<td>1.9</td>
</tr>
<tr>
<td>Industry</td>
<td>62</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Discussion
This study inevitably suffered from the varied quality of the individual contributions by our many recorders – see the acknowledgement below. We could check that the cited papers matched those in the stories because often these listed some of their details, such as the name of the journal or one of the institutes where the work was done, but we could not verify the extent of coverage, the details of the commentators or the accuracy of the codes that the recorders entered in their spreadsheets because of the languages of the stories. For some countries and newspapers, the amount of coverage did seem rather low, but one reason for this was that many stories gave few details of the research so that it was not possible to identify the paper(s) being cited. On the other hand, some newspapers were punctilious in giving their sources, notably Le Soir in Belgium and the Daily Mail in the UK.

We consider that our main conclusions are robust, as they appeared in the newspaper coverage in many different countries. These are that prevention and causation, rather than treatment, and that within treatment drugs rather than other therapies, are the main topics. It was also striking how nationalistic many journalists were, especially in the small countries as Figure 5 demonstrates. There was very little coverage of research from east Asia, even though most of it is in English and in international journals. This suggests a subtle racial bias which will cause readers to be unaware of the enormous increases in medical research output in these countries in recent years, and to continue to believe in the unchallenged supremacy of Europe and the USA in biomedical progress.

Acknowledgments
This project was supported by funding from the European Community’s Seventh Framework Programme under grant agreement EC/FP7/602536. The newspaper stories collection and identification of the cited research studies in the various European newspapers and languages was done by the following individuals, for Austria, Germany and Switzerland: Natalia Kelsch, Anne Spranger, Victor Stephani and Tobias Schumacher from Technische Universität Berlin, Germany; for Belgium: Ann-Sophie de Mol and Gabrielle Emanuel from King’s College London (KCL), UK; for Bulgaria: Eva Nacheva and Christina Tencheva from
KCL; for Croatia: Ria Ivandic Emanuel from KCL, UK; for Cyprus: Chryso Th. Pallari from the University of Cyprus, Nicosia, Cyprus; for the Czech Republic and Poland: Kasia Zemanek from KCL; for Denmark: Maria Dahl and Maria Emilsson from KCL; for Estonia: Argo Soon from the Estonian Research Council, Tartu, Estonia; for Greece: Laura Mantovani from KCL; for Hungary: Csajbok Edit from Semmelweis University, Budapest, Hungary; for Italy: Ludovica Borsoi from Università Commerciale Luigi Bocconi, Milan, Italy; for Latvia: Ingrid Jaselskyte, Estonian Research Council, Tartu, Estonia; for Luxembourg and the Netherlands: Ann-Sophie de Mol from KCL; for Portugal: Diana Gosałvez-Prados, Elisabeth María Ilidio-Paulo, Camila Higueras-Callejón and José Carlos Ruiz-Jiménez from Escuela Andaluza de Salud Publica, Granada, Spain; for Romania: Maria-Cristina Juverdeanu from KCL; for Spain: Diana Gosałvez-Prados and Elena Salamanca-Fernández from Escuela Andaluza de Salud Publica, Granada, Spain and Tahereh Dehdarirad from Universitat de Barcelona, Barcelona, Spain; for Sweden: Gustaf Nelhans from Sahlgrenska Göteborg, Göteborg, Sweden; for the UK: Argo Soon, Marleen Saidla and Tiina Tasa (Estonia) and Eva Nacheva from KCL partly assisted EP on the data collection. The Excel VBA programs were written by Dr Philip Roe of Evaluametrics Ltd, St Albans, England.

References

Fanelli D. (2013) Any publicity is better than none: newspaper coverage increases citations, in the UK more than in Italy. Scientometrics, 95(3), 1167-1177.


Main Path Identification involving Article’s Associated Attribute: A Case Study of Synthetic Biology

Wei Ling\textsuperscript{1,2}, Liu Chunjiang\textsuperscript{2}, Xu Haiyun\textsuperscript{2}, Chen Yunwei\textsuperscript{2}, Fang Shu\textsuperscript{2}

\textsuperscript{1}weiling@mail.las.ac.cn
School of Information and Management, Shanxi University of Finance and Economics, Taiyuan (China)

\textsuperscript{2}liucj@clas.ac.cn, xuh@clas.ac.cn, chenyw@clas.ac.cn, fangsh@clas.ac.cn
Chengdu Document and Information Center, Chinese Academy of Sciences, Chengdu China

Abstract
This paper proposes a new measure to trace the main paths of knowledge flows, which are characteristics of particular semantics and can cater to different research or application demands. The traditionally main path analysis usually neglects the inequivalence between citations (H&D, 1989), leading to inaccurate results. To address this problem, we take the documents’ associated attributes into account to measure the inequivalence, and further transform the inequivalence into a relevant relationship. Our method is a modification based on SPC from Batagelj (2003), using meta-path to describe and quantify the relevancy on the basis of associated attribute’s correlation, then incorporating the relevancy into the link’s traversal weight, and finally combining it with original SPC traversal weight to form a modified index. Synthetic biology is taken as a case study to test the performance of the new index and the results proved that our method can be an effective complement to main path analysis. Our method mainly boasts two beneficial effects: (1) it further developed the methodology of main path identification. A heterogeneous bibliometric network based on citations was constructed, and meta-paths were used to enrich main paths with more information and semantics; (2) the application scenario of main path analysis was expanded. Selected associated attributes can be integrated into main paths, catering to different research or application demands.

Keywords: main path analysis; SPC; associated attribute weight; modified traversal weight

Conference Topic
Methods and techniques
Citation and co-citation analysis
Mapping and visualization
**Introduction**

Citation network is the collection of citing and cited articles (Price, 1965), reflecting the knowledge transfer from cited articles to citing articles. With the passage of time and rapid growth of literatures, citation network has become a huge and intricate system, making it hard to figure out paths in a specific research field. However, identifying the disciplines’ development trajectory from the large and complex citation network is a necessity for researchers to understand and grasp the developing trends in their field, which fosters the practical demand for main path analysis. Meanwhile, research results prove that a discipline is mostly defined by a few important literatures and scholars in its history of development (Small, 1970), providing theoretic support for main path analysis. Main path analysis has not only been used in the article citation networks to identify core papers and knowledge diffusion paths, but also applied in the patent citation networks to identify key technologies and describe technologies’ evolution paths.

Specifically, the main object of this study is to propose a new measure; it can track the main paths of knowledge flows, which are characteristics of particular semantics, and caters to different research or applied demands. Our method is modified on the basis of SPC (Batagelj, 2003) in main path analysis (H&D, 1989). The main difference is that documents’ associated attribute is analyzed and employed in our method to analyze the inequivalence between the citations. With this new method, inequivalence can be transformed into a relevant relationship and meta-path of heterogeneous bibliographic network will be used to describe and quantify the relevancy based on associated attribute’s correlation. The relevancy was then added to link’s traversal weight.

We conducted a case study on synthetic biology to test the feasibility and applicability of our new method. The comparison between our index and SPC proved that our method can be an effective complement to MPA. The main contributions of our study can be summed up in three aspects: (1) enriching the methodology of main path identification; (2) expanding the application scenario of MPA; (3) revealing the knowledge diffusion path of synthetic biology domain. The rest of the paper is organized as follows. Section 2 introduces MPA and related works, followed by Section 3 which presents research questions and methodology. Section 4 is a case study, which applies our new method and original method to a set of papers related to synthetic biology, and a comparison is made between the results respectively obtained by ours and by the original one. Finally, we give the discussion and conclusion.

**MPA and Related work**

Hummon and Doreian (1989) put forward a method called “main path analysis” which uses citations’ similarity to identify the most representative articles and theories. The method offers three indices to exam connectivity in an acyclic network, which are NPPC (node pair projection count), SPLC (search path link count) and SPNP (search path node pair). The basic procedures of MPA can be summarized as follows: first, one of three indices is chosen to assign weights to all links in the network; then, starting at the source nodes, a greedy method is used to search a link sequence and this process is repeated until all the source nodes are touched; at last, the sequence that is the highest in sum of link weights is selected. However, our paper aims to introduce another index SPC, proposed by Batagelj (2003). In a citation network, when all paths from every source to each of the sinks are searched, the SPC for each link is defined as the total number of times the link is traversed. We choose SPC index as the basis and comparison object of our method.

According to Wang’s work (2013), we divide MPA into five parts: ① network topology. As a
directed network, the network topology order of citation network is fixed; ② links’ initial weights, which are equal to 1 in the form of traversal count; ③ links’ traversal weight index (SPC); ④ main path selection principle, generally selecting the main path with largest sum of traversal weights; ⑤ starting node selection of searching, namely where to begin the search in the network. Here are the milestone works on ④ main path selection principle. Aiming to address the problem of local optimum caused by greedy strategy, Verspagen (2007) extracted all paths from sources to sinks and chose the path with the largest sum of weights as main path. Liu (2012) suggested a method of multiple main paths which relaxed the search constraint to obtain paths with the top overall traversal counts and thus revealed more details. Bekkers (2012) proposed to extract the main paths according to H&D, and then add the nodes with knowledge contribution to main paths. Then, we move to the works on ⑤ starting point selection of searching. The original main path analysis is featured by searching forward, to be specific, searching from sources to sinks. Liu (2012) uncovered a new method, which is searching backward from the latest papers to the earliest papers. Liu also suggested “key-Route Search” to solve the potential problem that the link with the highest traversal count may be excluded.

we are most concerned about ② links’ initial weights and ③ links’ traversal weight index. Yeo & Kang (2014) applied the aggregative approach to improve SPC and other link count indices and used the stochastic approach to carry out path dependent search to separate merging papers. Zhu (2015) substituted an integrated weight for the traversal weight to identify main paths. The integrated weight was composed of an external correlation of citations such as citation, co-citation and coupling relationship and an internal text correlation between the citing paper and cited context. Chen (2016) translated patent documents into sets of vectors and calculated the semantic similarity between patents to search multiple main paths which focus on certain technology topics.

The above researchers have made great contributions to overcoming the limitations of MPA and expanding the concepts and unlocking characteristics of MPA, yet few researchers attach importance to citations’ inequivalence, which is important but unexplored. To address the problem and fill the void in the literature, this research will further probe into ③ links’ traversal weight index.

Research questions and new index

Research questions

So far, MPA either has treated citations equally or calculated the correlation between the citing and cited documents based on the text content. However, both ways revealed the structure of knowledge merely from the literature itself, and failed to take documents’ attached information such as authors, institutions and journal into account. We define this kind of attached information as documents’ associated attributes. According to our literature investigation result, few researches are carried out with the consideration of documents’ associated attributes. In the paper, three research questions are raised: (1) How to describe and quantify the relevancy between the citing and cited documents based on their associated attributes? (2) How to modify the traversal weight by using the relevancy and SPC index? (3) What are the differences between main paths identified by SPC traversal weight and our modified traversal weight? In order to answer the above questions, we tried a variety of methods including heterogeneous bibliographic network and meta-path, linear combination and comparative analysis.
Meta path for new weight

In a heterogeneous information network (HIN), two objects can be connected via different paths called meta paths (Sun & Hun, 2011). Meta paths are abundant in semantics and different meta paths imply different correlations or relationships between objects. Heterogeneous bibliographic networks, which capture the semantics of a real-world network (Sun & Hun, 2013), are a typical example of HINs. On the basis of citation relationships, we construct a heterogeneous bibliographic network in which nodes can be connected via different associated attributes to form various kinds of direct or indirect relationships. Meta paths that show objects’ direct relationships are short; however, this type of meta paths can also become long by concatenating them through direct relationships. Direct relationships (meta paths) extracted from the bibliographic data and two long meta paths are presented in Table 1.

Table 1 Meta paths in a heterogeneous bibliographic network

<table>
<thead>
<tr>
<th>Relation type</th>
<th>meta path</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct relation (short meta path)</td>
<td>P → P</td>
<td>Paper citing a paper</td>
</tr>
<tr>
<td></td>
<td>WB → w</td>
<td>Paper written by an author, an author writes paper</td>
</tr>
<tr>
<td></td>
<td>P → A, A → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WB → I, I → P</td>
<td></td>
</tr>
</tbody>
</table>

It is necessary to stress that the starting node and ending node on meta path are restricted to citing and cited documents. For example, two papers can be connected via “paper → author → paper → author → paper” path or “paper → institution → paper → institution → paper” path, which means two papers forms a citation relationship and meanwhile the relationship is elaborated with such connections as “author” and “institution”. The chains extend and progress with these connections which also write other similar paper(s). In other words, the citation relationship also involves co-author relationship or co-institution relationship.

Associated attribute weight (AAW)

We want to know whether and how documents’ co-author relationship or co-institution relationship relate to the citation nodes. We use the meta path to describe and qualify the relevant relationship, and define the relation as associated attribute weight (AAW), that is, documents’ similarity based on meta path. Considering that authors or institutions in citing and cited documents may have weak connections, a long symmetric meta path is constructed. Allowing for the meta path’s symmetry and nodes’ homogeneity, Pathsim (Han & Sun, 2011) is used to calculate nodes’ similarity.

When a symmetric meta path $\mathcal{P} = (A_1 A_2 \cdots A_n)$ is given, PathSim between two objects $x_i$ and $x_j$ of the same type can be calculated by $s(x_i, x_j) = \frac{2M_{ij}}{M_{ii} + M_{jj}}$. Formula (1), where $M$ is the commuting matrix for the meta path $\mathcal{P}$, $M_{ij}$ is the number of instances of the paths between $x_i$ and $x_j$.
$x_i$, $M_{ii}$ and $M_{jj}$ are the visibility for $x_i$ and $x_j$ respectively in the network. $M$ is defined as $M = W_{A_1A_2}W_{A_2A_3} \cdots W_{A_{i-1}A_i}$, where $W_{A_iA_j}$ is the adjacency matrix between type $A_i$ and $A_j$. For the long meta path, it's convenient to calculate the nodes' similarity by path concatenation. Suppose $\mathcal{P} = (\mathcal{P}_i\mathcal{P}_i^{-1})$, $M_p$ and $M_p^T$ are commuting matrices of $\mathcal{P}_i$ and $\mathcal{P}_i^{-1}$, and $\mathcal{P}$ 's commuting matrix is $M = M_pM_p^T$. The meta path can be obtained by repeatedly concatenating $\mathcal{P}$ for $k$ times is signified as $\mathcal{P}_k$ or $(\mathcal{P}_i\mathcal{P}_i^{-1})_k$, and its commuting matrix is $M_{pk} = M_k$ or $(M_p^k = M_pM_p^T)_k$ Formula (2).

Taking meta path $P \xrightarrow{WB} I \xrightarrow{W} P \xrightarrow{WB} I \xrightarrow{W} P$ for example, which is symmetric and can be written as $\mathcal{P} = (PIP I)$. It is obtain by concatenating short meta math $PIP$ for two times, whose commuting matrix is defined as $M = W_{p1}W_{ip}W_{p1}W_{ip} = (W_{p1}W_{ip})^2$. $W_{p1}$ and $W_{ip}$ are the co-occurrence matrix of papers and institutions, which are transposed each other. Thus, $M$ can be transformed into $M = (W_{p1}W_{ip}^T)^2$. Table 2 shows the sample commuting matrix of meta path $PIP$ between five papers. Table 3 is the corresponding path similarity calculated by formula (1), namely the AAW in the paper.

<table>
<thead>
<tr>
<th>Table 2 Sample commuting matrix</th>
<th>Table 3 Sample AAW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P1$</td>
</tr>
<tr>
<td>$P1$</td>
<td>8</td>
</tr>
<tr>
<td>$P2$</td>
<td>0</td>
</tr>
<tr>
<td>$P3$</td>
<td>4</td>
</tr>
<tr>
<td>$P4$</td>
<td>0</td>
</tr>
<tr>
<td>$P5$</td>
<td>0</td>
</tr>
<tr>
<td>$P1$</td>
<td>1</td>
</tr>
<tr>
<td>$P2$</td>
<td>0</td>
</tr>
<tr>
<td>$P3$</td>
<td>0.44</td>
</tr>
<tr>
<td>$P4$</td>
<td>0</td>
</tr>
<tr>
<td>$P5$</td>
<td>0</td>
</tr>
</tbody>
</table>

Modified transversal weight (MTW)

When MPA is applied to identify main paths, Pajek will automatically calculate all links’ SPCs according to the network’s topology. More clearly, this SPC refers to SPC traversal count. In order to smooth the range of traversal values, and eliminate the degree of variation, it is necessary to normalize the counts. After being normalized, SPC traversal count will be transformed into SPC traversal weight (SPC TW).

Next efforts are made to construct a new index based on SPC traversal weight. We make a combination between the associated attribute weight (AAW) and the original SPC traversal weight, and define it as modified transversal weight (MTW). According to the above analysis, two questions have to be answered. The first one is how to normalize the SPC traversal count. Three ways of normalization are included in Pajek, namely normalized by flow, max and logarithmic. Results acquired through these three normalization ways will be compared mutually, and compared to the associated attribute weights (AAWs). One normalization way will be chosen. The second research question is how to combine the two weights. Since there are so many zeroes in the AAWs, we have to make a linear combination, which is expressed in $MTW = \alpha SPCTW + \beta AAW$. Formula (3). In the formula, $\alpha$ and $\beta$ are weight coefficients, which will be fixed after the actual data was observed and tested.

Case study: Synthetic Biology

Data collection and processing

We retrieved data from the web of science database which has been widely used by worldwide researchers. It offers the citation information dating back to 1864. We retrieved 869 papers with the
query of “synthetic biology” to a topic field, and the time horizon was set from 1900 to 2010. It should be noted that if the range of retrieve time was set until 2016, over 7000 papers will be obtained, which will form an extraordinarily large pool of authors and institutions. As a result, the research topics will be quite disperse. To avoid this problem, the paper selects the synthetic biology literature from 1900 to 2010 to identify the main paths and analyze the relevancy between associated attributes and citation nodes. In this way, the research process becomes more feasible and the research purpose can be effectively fulfilled.

After duplicate titles were removed, papers without authors and institutions were deleted and the final number of papers amounted to 859. All the 859 records were then imported in CitNetExplorer to generate citation network within the scope of data set. The largest connected component with 636 nodes and 2607 edges was extracted as the research object.

Associated attribute “institution” was taken for example to compare main paths separately identified by SPC and our method. Firstly, SPC traversal weights (SPC TWs) were calculated and 2607 SPC traversal counts were obtained ranging from 1 to 72680. Results obtained from the three ways of normalization show that the two ways normalized by max and flow are not suitable and that the traversal weights are too small to combine with AAWs. Logarithmic weights are reasonable. 339 edges’ traversal weights are zero and we corrected them as 0.01 to ensure the network’s connectivity. Secondly, associated attribute weights (AAWs) were calculated. Meta path \( P = (PIP) \) was constructed and its commuting matrix was obtained by using Formula (2). AAWs of all citations were then calculated by using Formula (1). AAWs of 365 citations are nonzero, ranging from 0.027 to 1. At last, modified traversal weights were calculated (MTWs). Four groups of test value were set for \([\alpha, \beta]\), that is \([0.9, 0.1] , [0.8, 0.2] , [0.7, 0.3] \) and \([0.6, 0.4]\). It is found that the group \([0.7, 0.3]\) is reasonable by comparing SPC TWS, AAWs and MTWs. Finally, \( MTW = 0.7SPC\ TW + 0.3AAW \) Formula (4) was deduced. All links’ MTWs were calculated by Formula (4).

Experiments and results

We tested SPC AW and AAW on the citation network, identified two kinds of forward local main paths, multiple forward local main paths, backward local main paths, multiple backward main paths, standard global main paths and global key-route main paths, and made a comparison to verify our method’s effectiveness and necessity. Due to the space constraint, we will elaborate on the findings of the multiple local main paths, and make a brief introduction to the findings of the other main paths.

Results of SPC traversal weight-Multiple forward local main path

A total of 22 papers are found in the forward multiple main path shown in Fig. 1, including 17 papers of forward local main path (labeled in yellow) and the other 5 papers (labeled in blue, No 85,87,108,210,283). The thicker the line is, the higher SPC index is. The node is marked in the form of “serial number in the dataset-first author (publication year)”. The arrowhead indicates the cited article, and the tail indicated the citing article.

1-Hobom (1980) published the first paper titled as “synthetic biology” to describe a bacteria genetically engineered through recombinant DNA technology. Seven important papers sprung up in the year 2004 and 2005. 9-Hendrickson (2004) introduces in detail how various types of polymerase were widely differed in the interaction with c³dA. 10-Benner (2004) describes the roles of nucleobases and phosphates of DNA in the molecular recognition processes played by synthetic
chemistry. 23- Sismour (2004) is a follow-up study to the Benner’s. 8-Ball (2004) puts forward some questions about the risks of redesigning life and 18-Benner (2005) made a response. Two works of Sismour (2005) are further explorations on synthetic techniques.

Both 61-Chin (2006) and 107-Drubin (2007) are important review articles on advances in synthetic biology. 240-Filipovska (2008) focuses on “orthogonal” approach to build metabolism within the cell. 279-Purnick (2009) summarizes the research works before as the first wave which is featured by “cells-level” and the second wave on the way “systems-level”. 284-Carr (2009) affirms the trend by looking back on 50 years’ history. 295-Lu (2009), 454-Khalil (2010) and 455-Cullura (2010) are achievements of a joint library set up by several top biomedical research institutions. 295-Lu (2009) looked forward to the next generation synthetic gene networks, 454-Khalil (2010) reviewed the practical applications of synthetic biology, and 455-Cullura (2010) confirmed the advantages of the lab’s RNA-based system. 456-Nissim (2010) constructs a dual-promoter integrator to target diseased cells. It’s clear that the citation between 295-Lu (2009) and 454-Khalil (2010) had the largest weight, which plays an important role in the development in synthetic biology. The other 5 papers locate together, topics of which focus on the quantitative experiments and standardization measures of gene recombination.

Results of SPC traversal weight-Multiple backward local main path

34 papers were found in the multiple backward local main paths shown in Fig. 2, among which 23 papers were labelled yellow and formed the backward local main path.

In the year 2010, research on synthetic biology is in the stage of rapid development, with the number of academic works mushrooming. 454-Khalil (2010), cited by ten papers, is in the core of the whole path in terms of the practical applications of synthetic biology. The citing papers make a variety of further innovative explorations in different fields. For example, 488-Weber (2010) highlights the advances in mammalian synthetic biology while 628-Fritz (2010) pointes out two developing strategies and a conceptual model to improve the performance and the structure of biological systems.


In conclusion, forward local main path presents fundamental articles, important researchers and the knowledge diffusion traces between the articles and researchers; by contrast, backward local main path reveals the hot topics and research frontiers, and their backtracking of research basis or references. Multiple main paths can give more details.

Results of modified traversal weight-Multiple forward local path main

In spite of the relative complexity of the multiple forward local main paths (Fig. 3), it can still be clearly divided into two sub paths, including the forward local main path labeled in yellow. The forward local main path is relatively short with only 9 papers. It is quite different from the forward local main path obtained by SPC AWs, since they have nothing in common except the three head codes. Obviously, the AAWs play an important role in directing the main path.

Four citations have a large AAW. AAWs of citations between 9-Hendrickson (2004), 10-Benner (2004), and 18-Benner (2005) are equal to 1 because they come from the same institution, Department of Chemistry, University of Florida. 113-Marguet (2007), 301-Tanouchi (2009) and 346-Pai (2009) are connected by Department of Biochemistry and Institute for Genome Sciences and Policy, Duke University. 530-Armengaud (2010) and 531-Christie-oleza (2010) are colleagues of Lab Biochem System Perturb. On the whole, the forward local main path consists of three research groups, which successively reviewed the recent advances of synthetic biology along the time line.

The upper sub path and the forward local main path underneath are connected by 19-Medanial.
The upper sub path is comprised of four research groups. 19-Medanial (2005) and 13-Basu (2004) have the same corresponding author Ron Weiss, who comes from Department of Molecular Biology, Princeton University. Green branch, 16-Rackham (2005), 61-Chin (2006), 62-Chin (2006), shows the achievements of Medical Research Council Laboratory of Molecular Biology, in building new functions in living organisms. 47-Anderson (2006) and 85-Anderson (2007) are publications of Howard Hughes Medical Institute, Physical Biosciences Division, Lawrence Berkeley National Laboratory and other institutes, mainly focus on environmental information sensing and integration. 167-Lee (2008), 350-Carothers (2009) and 570-Carothers (2010) belong to the Joint Bioenergy Institute, which includes Sandia National Laboratory, Lawrence Berkeley National Laboratory, University of California and other branches.

Results of modified traversal weight-Multiple backward local main path

The multiple backward local main paths (Fig. 5) are much more complex and can still be divided into two sub paths, including the backward local main path labeled in yellow. The nodes labeled in blue without distinct index values won’t be stressed.

Fig. 5 Multiple backward local main path at 30% tolerance (38 papers)

On the backward local main path, the institutes’ correlation is much lower and only three citations have large AAWs. 3-Zhou (2004) and 4-Tian (2004) are important pioneers on microfluidic technology for multiplex gene synthesis. They are commonly connected to three institutes, namely Department of Chemistry, University of Houston, Atactic Technologies, Houston, and Chemical Engineering Department, University of Michigan. 21-Sprinzak (2005) and 109-Haseltine (2007) research synthetic gene circuits from different angles and they both belong to California institute of technology. 293-Gulati (2009) and 524-Szita (2010) publish two reviews about microfluidic approaches on a themed issue about Systems Biology and their co-institutes include University College London and Imperial College London. Papers of 19-McDaniel (2005) to 292-Boyle (2009) are related to synthetic gene circuits or micro-organisms combination techniques yet with much smaller AAWs. Institutes on the branch 110-Yeh (2007), 171-Lucks (2008), 421-Skerker (2009), and 484-Skar-gislinge (2010) seldom cooperate with each other.

On the below sub path, authors of blue branch 19-Medanial (2005), 106-Batt (2007), 279-Purnick (2009) come from Department of Electrical Engineering, Princeton University. Green branch 295-Lu (2009), 454-Khail (2010), 455-Callura (2010), 456-Nissim (2010) presents cooperation between Howard Hughes Medical Institute, Harvard University and Boston University. These three universities or institute have conducted many empirical experiments on synthetic bacteriophage or riboregulator. The upper sub path contains three research groups including 9-Hendrickson (2004) to 24-Sismour (2005), 167-Lee (2008) to 570-Carothers (2010), and 107-Drubin (2007) to
Comparison between results

This section makes a comparison between the main paths separately identified by SPC traversal weight (generally called as SPC main paths) and modified traversal weight (generally called as MTW main paths). The similarities and dissimilarities on the path structures, path changing trend, path functions and application scenarios are summed up as following:

Path structures and changing trend. Contrary to SPC local main paths which are longer with a number of starting nodes and a single ending node, the modified local main paths are shorter with one single starting node and ending node, and can directly and clearly represent the citations with the maximum relevance of institutes.

SPC multiple main paths are somewhat disordered and unseparated, and present the most representative articles and theories. While modified multiple main paths with clear skeleton are much more regular and can be separated, which describe multiple knowledge flows. Besides, two kinds of modified multiple main paths show great differences. The multiple forward main path begins from the earliest time with “a narrow entrance” due to fewer literatures. It reveals the most important sources of the research domain, and new research groups’ emergence, development and cooperation following the sources. On the opposite, the multiple backward main path starts from the most recent time with “a broad entrance” because of the large amount of literatures. It can pick up multiple topics and trace their historic origins, so it’s relatively theme-focused. Along the main path, the earlier research works and groups related to the particular theme can be found.

Two global main paths are almost identical. It is simply because the AAWs only accounts for 30% of the modified traversal weight and many of AAWs are zero. However, the modified global main path can still reveal the institutes’ connectivity between the citations. Local key-route main paths are sensitive to the parameter and specifically a small increase in parameter will cause a large increase in the number of nodes. In addition, two kinds of main paths cannot be easily compared owning to their great complexity and difference. A small-scale global key-route main paths will have identical nodes, but the paths become more diverse as the scale is enlarged. Besides, the trend of modified main paths shows more rapid and notable growth.

Functions and application scenario. SPC main paths present the the most representative articles and theories according to the citation network’s topology and connectivity. While MTW main paths reveal those important articles with specific characteristics due to network connectivity and associated attribute, present the process and history of knowledge diffusion in a view different from SPC main paths, supplying more detailed information in a particular scenario.

Discussion and Conclusion

The aim of this paper is to find a new method of main path identification by constructing an index on the basis of SPC traversal weight. We have quantified the relevance between the associated attribute and citation nodes, and integrated it into weight to identify special main paths.

Based on the results from the case study, the conclusion can be drawn that our method has the following beneficial effects: (1) methodology of main path identification is developed and improved. A heterogeneous bibliometric network is built based on citations, and meta-path is used
to enrich the information and semantics of main paths. We made an investigation on the citing and cited paper pairs connected via $P \rightarrow I \rightarrow P \rightarrow I \rightarrow P$ to find out whether and how institution may affect the main path. Research results showed that, 365 out of 2607 citations embedded an institution co-cooperation relationship, and such semantic relationship was well presented by the modified main paths; (2) application scenario of MPA is expanded. Selected associated attribute can be integrated into main paths, catering to different research demands. The example “institution” successfully demonstrates that main paths can unveil important research institutes, outstanding groups and their cooperation. For example, in the early stage, Steven A. Benner and A. Michael Sismour of Department of Chemistry, University of Florida, and Cynthia L. Hendrickson of Nuclear Medicine Department made much contributions to synthetic biology’s development. There are more internal cooperation. In the subsequent period, some research institutes gradually stood out, such as Department of Biochemistry and Institute for Genome Sciences and Policy, Duke University, Department of Molecular Biology, Princeton University, Howard Hughes Medical Institute, Lawrence Berkeley National Laboratory and etc. Scientific cooperation tends to be external and cross-organized. We consider the modified main path analysis method can be used to help to do agency evaluation; (3) the knowledge diffusion path of synthetic biology domain is identified to help researchers know about the domain quickly.

This study makes an initial contribution to main path identification concerning the documents’ associated attribute and shows an effective performance; however, the limitations should be considered when interpreting the results. First, our method is tested on synthetic biology, other academic fields are excluded due to limited time. Second, the study is restricted to testing a single attribute, failing to unlock the full potential of this method. Thus, further studies on different attributes and academic fields are needed to improve this method. The future researches can be done in the following directions: (1) choosing and analyzing other associated attributes and integrating them into the traversal weight for new indices; (2) trying asymmetric meta paths with different lengths to describe more kinds of relevancy; (3) comparing the results in this paper with those possible ones obtained through (1) and (2) for further findings.

Acknowledgements
This work was supported by the China Postdoctoral Science Foundation Funded Project 2016M590124 and the West Light Foundation of the Chinese Academy of Sciences, China under grant no. [2013]165(3-6).

Main References


The effect of “open access” on journal impact factors: A causal analysis of medical journals

Cui Huang 1, Xiaoxu Yue 2, Jing Chen 3, Weixiao Xu 4, Jiang Li 5*

1 hcui@tsinghua.edu.cn
Tsinghua University (China)
2 Yuexx2013@163.com
Tsinghua University (China)
3 Chenjing24@mail.tsinghua.edu.cn
Tsinghua University (China)
4 xuwx07@163.com
Zhejiang University (China)
5 li-jiang@zju.edu.cn
Zhejiang University (China)
* Corresponding author

Abstract
The Journal Impact Factor has a significant influence on author of submission proverbially. On the basis of a panel data analysis method, this present study aimed to clarify whether open access (OA) is beneficial to the Journal Impact Factors or the dissemination of academic achievements. The study was carried out in analyzing journals of medicine in databases such as Web of Science and Ulrich’s Periodicals Directory. The results indicate a causal relationship between the journal impact factors (JIFs) and OA, specifically that: (1) OA enhances JIFs; (2) countries that are less developed in science and technology are more likely to choose OA; (3) control variables such as language, amount of published papers, and release cycles all have a significant impact on JIFs.

Conference Topic
Journals, databases and electronic publications
Introduction

With the continuous development of information technology, social environment, and network technology, the traditional ways to disseminate academic achievements have been changed. Undoubtedly, the digitalization of research articles and the establishment of online academic communities will have a certain impact on the pre-publication stage. In addition, modern journal publications are facing problems such as high prices, publication delay, and difficulties in public acceptance. These issues have prompted publishers and researchers to seek corresponding solutions. The establishment of online network information exchange platforms for academic papers not only can effectively save resources, but also can help researchers understand the latest academic trends and reveal research frontiers in a timely manner. Therefore, open access (OA) came into being in the late 1990s. By the end of February 2016, the number of journals included in the world’s two largest OA journal retrieval platforms (i.e., Directory of Open Access Journals and Open J-Gate) increased by 1.5 times (from 7,522 to 11,441 journals) and 2.7 times (from 8,300 to 22,000), respectively, when compared with the same period in 2012. These two platforms also show that the number of OA papers has increased by 2.9 times (from 761,000 to 2,233,000) and 3.9 times (from 2,000,000 to 7,840,000) during those five years (2012-2016).

Literature review

The idea of impact factor was first proposed by Eugene Garfield (1955) in Science. Garfield and Sher (1963) introduced Journal Impact Factor (JIF) as a tool to select journals for the Science Citation Index (SCI). Since 1976, JIFs, clearly using the formula still used today, are published on a yearly basis in the Journal Citation Reports (Shi et al., 2017). Currently, JIFs have been widely used as an index to measure the quality of journals and, even worse, articles published in it (Alberts, 2013; Bollen, Rodriguez, & Van de Sompel, 2006; Kurmis, 2003; Seglen, 1997; Simons, 2008; Vanclay, 2009).

The previous research on the influence of OA papers mainly investigated the advantages of OA citations, namely whether OA journals were more easily cited than non-OA journals. The main conclusions from this literature review were: 1) OA papers have had certain advantages in citations, but there were obvious differences of citation advantage among different disciplines (Antelman, 2004; Craig et al., 2007; Davis, 2009; Eysenbach, 2005; Gaulé, 2009; Hajjem et al., 2006; Metcalf, 2006; Norris et al., 2008; Piwowar et al., 2007; Schwarz, 2004; Tonta et al., 2007). Researchers have found that OA can increase citation frequency, but there are varying degrees of impact on different disciplines. 2) No citation advantages for OA papers. In 2008, Davis et al. (2008) studied the relationship between OA and citation frequency based on random trials. The results showed that OA journals downloads were more than non-OA journals in the first year of publication, but none was highly cited. Moed (2007) argued that the citation advantage was related to quality bias and early view effect, and therefore had no relevance as to whether the paper was OA or not. Lansingsh (2009), using a multiple linear regression model and controlling the characteristics of papers, did not find a significant relationship between the paper acceptance and its citations at that period in time. Calver (2010) adopted a linear regression model to analyze eight journals of biology found that the OA citation advantage has disappeared after controlling for the characteristics of papers and increasing control for the characteristics of the authors. Based on a quantile quadrant analysis, Koler-Povh (2014) analyzed 2,007 papers in OA journals in the field of civil engineering, and the results also showed that there was no significant relationship between OA and citation frequency. Given the lack of consensus in the literature, does OA really promote the spread of academic achievements and increase JIFs? Is there a certain correlation between this effect and other factors?

In previous studies, the influence of OA in disseminating academic achievements was mainly explored through descriptive comparisons, random tests, and regression analyses. There are far fewer studies on causal relationships
and most used only one or a few influencing factors that are not systemic. In addition, most previous studies focused on the effect of OA on the number of long-term citations, rather than short-term citations, e.g., the citation window considered in JIFs. Therefore, based on general influencing factors for academic achievements, a quantitative method from the field of social science was adopted to systematically investigate causal inferences and to explore the relationship between OA and the impact factors of academic journals. The implications of this analysis include suggestions for improving academic exchange mechanisms, constructing academic exchange platforms, and transforming academic achievements.

**Data and methodology**

**Data source**

It is worth drawing attention to some of the advantages of selecting the field of medicine for our study. Firstly, medical journal publications are characterized by long spans of time and a relatively complete system. Since *New Advances in Medicine* began its publications in France in 1665, many varieties of medical journals ranging from comprehensive content to specialized subjects have been developed in the last 300 years, which means that there is now a robust system of academic publications in the field of medicine. Secondly, there is a wide distribution of medical journal publications in most countries around the world. Specifically, our statistical analysis showed that as of the end of February 2016, at least 84 countries have medical journal publications that correspond to 36 different subfields within medicine, based on the Web of Science (WoS) subject categories. Thirdly, it is feasible to obtain a relatively complete sample of publications from various databases. Many publishers have established comprehensive medical databases, such as the Medline, the PubMed, the WoS, etc. Therefore, sampling medical journal data implies having a reliable representation of the population, as well as a high level of data collection feasibility.

The data in this study consist of two parts. The first part came from the Journal Citation Report (JCR) published on the Web of Knowledge, of which there are 36 subjects related to medicine out of a total of 226 catalogued subjects. The data we collected from JCR fully included 36 subjects from the 2011-2015 period, based on total citation frequency between 2010 and 2014, JIFs, and impact factors with self-citations removed. In the 2015 JCR, a total of 2,404 journals were screened from the 36 subjects in the field of medicine. 2,045 journals were left after excluding duplicates and invalid JIFs. The remaining dataset included 172 OA journals and 1,873 non-OA journals. OA journals were identified as such in the JCR.

The second part of the data came from Ulrich's Periodicals Directory, the websites of journal publications, ResearchGate, Google Scholar, etc. From these platforms, the information was collected concerning: changes in the editor, changes in the name of the journal, the year of the first publication, OA year, academic subject information, number of papers, release cycle, language, country of publication, publishing medium, publisher, and other pertinent data.

**Methodology**

In the field of social science, in order to prove causality, we encounter the problem of “endogeneity””. Because it is hard to know whether there is a disturbance factor, it can influence the cause as well as the effect. In this case, the results obtained by regression analyses using the ordinary least squares model (OLS model) may produce errors, which will no longer give us causal inferences (Chen, 2012). Although panel data analysis can solve individual heterogeneity to a certain extent, the regression model itself may contain endogenous explanatory variables. In this study, there may also be an endogenous problem between the OA and the JIF: is the OA leading to a change in the journal’s impact factor, or is it because of the change in the OA journals’ impact indicators and the choice of being OA? Therefore the common instrumental variable method was adopted to solve the endogeneity problem. Because
the OA cannot appear in the year or even second year after the selection of the journal, eliminating the interference of time factors must be considered, thus presenting a need to introduce a fixed effect model. Therefore, to avoid the problem of possible endogeneity between independent variable and dependent variable, the interference factors of time was eliminated. A first order lag fixed panel analysis model was combined with the fixed panel instrumental variables model to explore the causal relationship between OA and JIF in this study.

Analysis Model

Factor analysis

Based on previous studies as shown in Table 1, the impact factors that influence journals can be divided into two aspects: 1) internal factors, such as author group, number of published articles, release cycle, editor, language, journal type, journal name change, etc.; 2) external factors, such as subject, publisher, search platform, etc.

<table>
<thead>
<tr>
<th>Author</th>
<th>Data</th>
<th>Method</th>
<th>Factors influencing JIFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mabel (2007)</td>
<td>7 medical journals of WOS (1999-2004)</td>
<td>Regression analysis</td>
<td>The change of editor had a significant impact on JIF</td>
</tr>
<tr>
<td>Mueller (2006)</td>
<td>Internal medicine journals of WOS in 2003</td>
<td>logistic regression</td>
<td>JIFs was closely related to the language of the journal, the state of journal had little influence on JIFs</td>
</tr>
<tr>
<td>Jiang (2007)</td>
<td>Core journals of 9 disciplines in A guide to the core journals of China (2004 version)</td>
<td>Correlation analysis</td>
<td>H index increased, the impact factors also increased generally</td>
</tr>
<tr>
<td>Yu &amp; Ma (2008)</td>
<td>Panel data of 9 journals from 2004 to 2006</td>
<td>Multiple regression model</td>
<td>The ratio of funded articles, paper length, the power and influence of the organizer, types of journals and authors of paper had a significant influence on JIFs</td>
</tr>
<tr>
<td>Li Ping Yu (2010)</td>
<td>537 medical journals of Science and Technology Journal Citation Report of China in 2008</td>
<td>Quantile regression model</td>
<td>Overseas papers were irrelevant to influence factors; journal with high impact factors and low impact factors, and the average number of citations were irrelevant to JIFs; journal with high impact factors, the ratio of the fund articles was irrelevant to the impact factors</td>
</tr>
</tbody>
</table>

In addition, there were two main factors that affected the size of impact factor based on the definition of JIFs: one was the number of papers published in the journal, and the other was the citation frequency of the journal articles. Therefore, the factors influencing JIFs could also be considered from the perspective of number of papers and citation frequency.

Variables

Based on the previous section, the impact factors of domestic and international journals were summarized. The release cycle, the number of papers published in the journal, the year of first publication, the language, country of publication, and so on were considered to have an impact on JIFs. The selection of variables is shown in Table 2. The samples in our study were SCI journals from the field of medicine. We selected JIFs for a five-year period (2010-2014). Ensuring that each journal had an impact factor, 7,815 samples were selected in the end. Tables 2 and 3 show that the average
JIF of the sample was 3.058666; after removing self-citations, the average JIF was 2.806097. Only 4.4% of 7,815 journals chose OA. In other words, the vast majority (95.6%) of journals did not choose OA. In addition, the average age of journal history was 40.7 years. Journal history was at least 2 years, with the longest being 194 years. The average release cycle was 9.4 publications per year; the lowest issue was only one per year, and the highest was 60 (not considering publications that are issued irregularly). There was an average of 299.9 published articles in each statistical year corresponding to the JIF. The lowest number of articles in a journal was one per year, and the highest number was 3,458. Since dummy variables, such as country and language, were numerous, they were not included in Table 2.

Table 2. Characteristics of the main variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Definitions</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IF</td>
<td>Journal impact fact or</td>
<td>The JIFs from 2010 to 2014</td>
</tr>
<tr>
<td>IF_Self</td>
<td>Journal impact fact or (self-citations removed)</td>
<td>The JIFs (self-citations removed) from 2010 to 2014</td>
</tr>
<tr>
<td><strong>Independent variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OA</td>
<td>Whether OA or not</td>
<td>0-Not, 1-Yes</td>
</tr>
<tr>
<td><strong>Control variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>Journal history</td>
<td>The JIFs year-the year of start publication</td>
</tr>
<tr>
<td>Country</td>
<td>Publication country</td>
<td>Dummy variable</td>
</tr>
<tr>
<td>Issues_per_year</td>
<td>Release cycle</td>
<td></td>
</tr>
<tr>
<td>Articles</td>
<td>The number of articles</td>
<td>The number of articles in previous two years</td>
</tr>
<tr>
<td>Year</td>
<td>The year of JCRIF</td>
<td>Dummy variable</td>
</tr>
<tr>
<td>Languages</td>
<td>Publication language</td>
<td>Dummy variable</td>
</tr>
<tr>
<td><strong>Instrumental variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publisher</td>
<td>Publisher’s OA bias</td>
<td>The number of OA journals/ the number of all journals</td>
</tr>
</tbody>
</table>

Table 3. Descriptive statistics of the main variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Sample size</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Minimum value</th>
<th>Maximum value</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF (self-citations removed)</td>
<td>7,815</td>
<td>2.806097</td>
<td>4.872357</td>
<td>0.019</td>
<td>162.182</td>
</tr>
<tr>
<td>IF</td>
<td>7,815</td>
<td>3.058566</td>
<td>4.936694</td>
<td>0.043</td>
<td>162.5</td>
</tr>
<tr>
<td>OA or not</td>
<td>7,815</td>
<td>0.044274</td>
<td>0.205716</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Journal history</td>
<td>7,815</td>
<td>40.72681</td>
<td>29.58302</td>
<td>2</td>
<td>194</td>
</tr>
</tbody>
</table>
Analysis model construction

We first changed the regression model to address the issue of missing variables, and then used the two-stage least squares method. The specific models are as follows:

\[
JCRIF = \alpha + \beta_1 \text{OA} + \beta_2 \text{Age} + \beta_3 \text{Issue} + \beta_4 \text{Articles} + \beta_5 \text{language} \\
\quad + \beta_6 \text{Country} + \mu_i + \varepsilon_i 
\]  

(1)

\[
\text{JIFnscite} = \alpha + \beta_1 \text{OA} + \beta_2 \text{Age} + \beta_3 \text{Issue} + \beta_4 \text{Articles} + \beta_5 \text{language} \\
\quad + \beta_6 \text{Country} + \mu_i + \varepsilon_i 
\]  

(2)

In these models, \( \alpha \) represents constant term, “OA” represents whether the journal is OA, “Issue” represents the release cycle, “Articles” represents circulation, “language” represents language that the journal is published in, “Country” represents the country of publication, \( \mu_i \) represents the year fixed effect, and \( \varepsilon_i \) represents random interference.

Then, to ensure the reliability of the model, it was necessary to deal with the potential endogeneity problem. If the endogenous problem did not exist, the two-stage method of least squares would not be as effective as the OLS method. For this purpose, the Durbin-Wu-Hausman test (DWH test) was used to test endogeneity. The model is as follows:

\[
JCRIF = \alpha + \beta_1 \text{OA} + \beta_2 \text{Age} + \beta_3 \text{Issue} + \beta_4 \text{Articles} \quad + \beta_5 \text{language} \\
\quad + \beta_6 \text{Country} + \rho + \delta_i + \mu_i + \varepsilon_i 
\]  

(3)

In this model, “\( \nu \)” represents the residual of one-stage regression, and the t-test was performed on the original hypothesis (\( H_0: \rho=0 \)) by using cluster robust standard deviation. If the original hypothesis is rejected, then an endogenous problem would be considered to exist. Otherwise, the population structure variables were exogenous. The p-value of the t-test of the original hypothesis was 0.0001, and was therefore rejected at the 0.001 significance level. Therefore, it is necessary to introduce instrumental variables to estimate whether the variable “OA” is in the JIF.

In order to deal with the potential endogeneity problem, and then accurately identify the causal relationship between OA and journal impact factors, this study took the publisher’s OA bias as the instrumental variable, and then used the panel data to produce a regression analysis of the instrumental variable. The equation for this is as follows:

\[
\text{OA} = \alpha + \beta_{\text{publisher}} + \gamma X_i + \delta_i + \mu_i + \varepsilon_i 
\]  

(4)

In this equation, “publisher” represents the OA bias of the publishing company. Publisher’s OA bias was highly correlated with the journal’s choice in having or not having an OA strategy. The OA status of the journal was entirely determined by the publisher. For instance, Biomed Central Company was a publishing firm dedicated to the promotion of OA that had a strong OA intention; all of its journals were therefore OA. In contrast, a total of 22 journals were

<table>
<thead>
<tr>
<th>Release cycle</th>
<th>7,815</th>
<th>9.410109</th>
<th>5.688468</th>
<th>1</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of papers</td>
<td>7,815</td>
<td>299.9419</td>
<td>290.408</td>
<td>1</td>
<td>3458</td>
</tr>
</tbody>
</table>

Note: Regression analysis included numerous dummy variables, such as country and language, which are not included here.
published by Elsevier Japan KK, none of which were OA. Elsevier Japan KK would therefore be labeled as “no OA intention”. And the publisher’s OA bias was used as a tool variable because there was no positive connection between the publisher’s OA bias and JIF. The weak instrumental variable test showed that the F-value was greater than 10, and there were no weak instrumental variables in the regression. The publisher’s OA bias could have been used as instrumental variables in this study.

Robustness test

Robustness testing is a necessary step to verify the suitability of the research model and the reliability of the results. To guarantee the reliability of the conclusions, we take into consideration that the influence of OA on the JIFs may be delayed. The panel data instrument variables model was thus used to verify the robustness of the model. Both the panel data and fixed effects models were able to elimination time deviations, but each had its advantages. The original model could be considered reliable if the results of the two methods were consistent.

Based on the regression results (Table 4), there were significant positive relationship between OA and JIFs, between JIFs and the release cycle, and between JIFs and the number of paper. However, the relationship between JIFs and journal’s year of first publication was still positive but not yet significant. That is to say, the robustness test results were consistent with the original model. Therefore, OA was considered to be one of the reasons that lead to the increase of JIFs in the field of medicine.

Table 4. Robustness test results

<table>
<thead>
<tr>
<th>Variables</th>
<th>Model 7</th>
<th>Model 8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>JIFs</td>
<td>JIFs (self-citations removed)</td>
</tr>
<tr>
<td>OA or not</td>
<td>1.836** (0.757)</td>
<td>1.820** (0.747)</td>
</tr>
<tr>
<td>Journal history</td>
<td>0.00456 (0.00633)</td>
<td>0.0051 (0.00627)</td>
</tr>
<tr>
<td>Release cycle</td>
<td>0.272*** (0.067)</td>
<td>0.263*** (0.0659)</td>
</tr>
<tr>
<td>Number of papers</td>
<td>0.00188** (0.00086)</td>
<td>0.00203** (0.00085)</td>
</tr>
<tr>
<td>Constant term</td>
<td>6.585*** (1.491)</td>
<td>6.347*** (1.468)</td>
</tr>
<tr>
<td>r2_a</td>
<td>0.0638</td>
<td>0.0591</td>
</tr>
<tr>
<td>rss</td>
<td>120430</td>
<td>118254</td>
</tr>
<tr>
<td>mss</td>
<td>10042</td>
<td>9228</td>
</tr>
<tr>
<td>rmse</td>
<td>5.104</td>
<td>5.058</td>
</tr>
<tr>
<td>r2</td>
<td>0.077</td>
<td>0.0724</td>
</tr>
<tr>
<td>F</td>
<td>1.9E+306</td>
<td>1.9E+306</td>
</tr>
</tbody>
</table>

Note:
1. The dummy variables year, country, and language were not identified in the table.
2. The regression coefficients of variables were tested by the z-test (normal distribution test) to examine the significance of regression coefficient.
3. *** p<0.01, ** p<0.05, * p<0.1

Results

Based on the regression analysis results shown in Table 5, all regressions were controlled for the fixed effects of time
Model 1 and Model 2 respectively represent the relationship between OA and JIF of non-self-cited journals when there were no control variables and instrumental variables. Model 3 and Model 4 respectively represent the relationship between OA and JIF of non-self-cited journals when instrumental variables were not introduced. Model 3 and Model 4 respectively represent the relationship between OA and JIFs of self-cited journal by adding instrumental variables. Regression analysis showed that the F-test was significant at the 0.01 level, and proved that the whole model was more significant and the regression results were more accurate.

From the regression results of Model 1 and Model 2, the journal hindered the promotion of JIFs and self-citations removed JIFs after the selection of OA but was not a significant effect. After adding the control variables, the regression results (Model 3 and Model 4) still showed that OA would hinder the promotion of JIFs and the self-citations removed JIFs. The parameter estimation of OA was smaller than no control variables. There were many reasons for the JIFs; both known and unknown impact factors would certainly have had an effect on the regression results. It was a negative impact but was not significant based on regression results. In other words, it did not have the explanation from the perspective of measurement. Therefore, the estimated size needed to be compared after adding instrumental variables.

Publisher’s OA bias can directly affect whether the journal chooses OA or not, but publishers cannot directly influence the JIFs, so instrumental variables were introduced to solve the endogeneity problem. The original regression results (Model 3 and Model 4) and regression results of Model 5 and Model 6 were observed after adding instrumental variables, which showed a reversal of the parameter estimates from negative to positive, and the results were significant. It was revealed that OA could increase the JIFs after removing endogeneity. The null hypothesis could therefore be rejected: OA had a significant influence on the impact factor of medical journals. OA was the cause of the JIF increase, and the OA could improve the influence of the journal.

Table 5. An empirical study of OA’s influence on JIFs

<table>
<thead>
<tr>
<th>Variables</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
<th>Model 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>JIFs</td>
<td>JIFs</td>
<td>JIFs</td>
<td>JIFs</td>
<td>JIFs</td>
<td>JIFs</td>
<td>JIFs</td>
</tr>
<tr>
<td>Whether OA or not</td>
<td>-0.326</td>
<td>-0.406</td>
<td>-0.469</td>
<td>-0.549</td>
<td>1.831**</td>
<td>1.795**</td>
</tr>
<tr>
<td>(0.709)</td>
<td>(0.699)</td>
<td>(0.706)</td>
<td>(0.696)</td>
<td>(0.753)</td>
<td>(0.742)</td>
<td></td>
</tr>
<tr>
<td>Journal history</td>
<td>0.0527***</td>
<td>0.0562***</td>
<td>0.00435</td>
<td>0.00489</td>
<td>0.00609</td>
<td></td>
</tr>
<tr>
<td>(0.00968)</td>
<td>(0.00955)</td>
<td>(0.00615)</td>
<td>(0.0066)</td>
<td>(0.00609)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Release cycle</td>
<td></td>
<td></td>
<td>0.268***</td>
<td>0.259***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.0671)</td>
<td>(0.066)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of papers</td>
<td>0.00104***</td>
<td>0.000844***</td>
<td>0.00182**</td>
<td>0.00198*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.00021)</td>
<td>(0.00021)</td>
<td>(0.00087)</td>
<td>(0.00086)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant term</td>
<td>3.097***</td>
<td>2.847***</td>
<td>1.245***</td>
<td>0.791**</td>
<td>5.955***</td>
<td>5.641***</td>
</tr>
<tr>
<td>(0.033)</td>
<td>(0.0325)</td>
<td>(0.397)</td>
<td>(0.391)</td>
<td>(1.378)</td>
<td>(1.354)</td>
<td></td>
</tr>
<tr>
<td>r2_a</td>
<td>0.333</td>
<td>0.333</td>
<td>0.32</td>
<td>0.321</td>
<td>0.071</td>
<td>0.0661</td>
</tr>
<tr>
<td>rss</td>
<td>3.537</td>
<td>3346</td>
<td>3501</td>
<td>3402</td>
<td>148302</td>
<td>145421</td>
</tr>
<tr>
<td>mss</td>
<td>0</td>
<td>0.247</td>
<td>36.04</td>
<td>33.68</td>
<td>13066</td>
<td>11979</td>
</tr>
<tr>
<td>rmse</td>
<td>0.869</td>
<td>0.856</td>
<td>0.864</td>
<td>0.852</td>
<td>4.897</td>
<td>4.849</td>
</tr>
<tr>
<td>r2</td>
<td>0.000045</td>
<td>7.18E05</td>
<td>0.0102</td>
<td>0.0098</td>
<td>0.081</td>
<td>0.0761</td>
</tr>
<tr>
<td>F</td>
<td>0.211***</td>
<td>0.337***</td>
<td>16.08***</td>
<td>15.46***</td>
<td>1.8E+306***</td>
<td>1.8E+306*</td>
</tr>
</tbody>
</table>

Note:
1. The dummy variables of year, country, and language are not identified in the table.
2. Standard deviation of regression variables are indicated in the parentheses.
3. The regression coefficients of variables was tested by the z-test (normal distribution test) to examine the significance of regression coefficient.
4. *** p<0.01, ** p<0.05, * p<0.1

The control variables include journal history, release cycle, number of papers, the country, the language and so on. “Journal history” refers to the total of years since the founding of the journal. For example, A Cancer Journal for Clinicians was first published in 1950, so its journal history is 66 years. There was a positive correlation between the JIFs and journal history but it was not significant, which showed that more established journals did not necessarily have higher JIFs.

The number of papers is a control variable directly related to the formula of the JIFs and has a close relationship to JIFs. The results showed that larger number of papers resulted in higher impact factors, while shorter release cycle also resulted in higher impact factors. However, there was a bigger effect of release cycle on the JIFs. The reason probably lies in the idea that researchers are more likely to submit papers to a journal with a larger number of published papers and shorter release cycle, therefore reinforcing this trend. This is obviously more beneficial to the rapid publication of scientific achievements.

The regression results were not displayed in Table 5 due to the excessive number of two dummy variables “country” and “language”. On the whole, the coefficients of country and language were both positive and significant. There could be a phenomenon related to the relationship between the country in which journals are published, their OA status and the JIFs: countries such as France, Australia, Germany and Sweden have had a higher average impact factors and a stronger overall influence on scientific development. The number of OA journals as a proportion of the total number of journals were lower in these countries, which also correspond to countries with relatively lower OA biases. Countries with lower impact factors and higher numbers of OA journals were more likely to have an OA bias, such as Columbia, Serbia, Uganda and other countries/regions. The analysis of countries with relatively weak scientific development influence revealed that most of journals were included in the WOS platform after selecting the OA, for example, Colombia Medica in Columbia. In addition, there was a positive relationship between the JIFs and language. English is the international language of scientific publications and therefore has the greatest impact on the JIFs among all languages. Most countries with smaller research capabilities have chosen English as the standard language of their academic journals. It could therefore be seen that OA was also an opportunity for these countries, not only to circulate scientific research achievements via OA platforms, but also gain free access to the most up-to-date international research data and results, thus promoting the development of science in a virtuous cycle.

Discussion and conclusions

In order to further understand whether the open access online environment could improve the impact of academic achievements, general instrumental variables and fixed effects models that were used to solve endogeneity problems in econometrics were adopted to study the causal relationship between OA and academic influence. The factors that may influence academic achievements were identified on the basis of previous studies and were treated as the control variables in our study. Our main research findings are the following:

First, OA can enhance the impact of academic achievement in the field of medicine. There are two main reasons behind this: One is free access and consequently, greatly increasing diffusion of the scope of knowledge; another is timely quality control based on peer review.

Second, countries with relatively low science and technology development are more likely to have an OA bias. France, Australia, Germany, Sweden and other countries have had a higher average impact factors, thereby also having a strong influence on scientific publications, and lower willingness to choose OA in the field of medicine, especially
because the top journals have relatively stable readership and reference groups. On the other hand, other countries such as Columbia, Serbia, and Uganda were more likely to choose OA with lower JIFs, which then enhanced the influence of JIFs to the point where they could be included in the WOS platform.

Last but not least, number of papers, release cycle, and the language of the publication also had a significant impact on the JIFs. English was found to expand the scope of knowledge diffusion because it is the international lingua franca. The larger number of papers and the shorter release cycle, the higher the JIFs.

Based on the results of the above analysis, it is suggested promoting the OA strategy for enhancing the influence of scientific research in countries with catch-up potential in science. However, since only high quality academic journal papers can be used to serve the majority of researchers, strict quality control procedures must be the prerequisite.

Some of the shortcomings in this study include: 1) Some journals may lack several JIFs due to being included in the SCI and then deleted, so the integrity of the entire medical field academic publication data cannot be guaranteed for a small number of journals. 2) The results of this study were limited to the medical field, so the conclusions may not be applicable to other disciplines. Future studies should add other academic fields for comparisons. 3) Insufficient sources of data. Citation data from SCOPUS and Google Scholar could be used in future studies to evaluate the influence of academic achievement in the OA environment.

Acknowledgements

We gratefully acknowledge the financial support of the National Natural Science Foundation of China (Grant Numbers 71673164, 71233005, 71673242 and 71520107005).

References

downloads, and citations: randomised controlled trial. *Bmj*, 337(7665), 1-6.


Analysis of highly cited papers in Rheumatology

Veronica Perez-Cabezas¹ Carmen Ruiz-Molinero¹ Ines Carmona-Barrientos¹
Enrique Herrera-Viedma² Manuel J. Cobo³ Jose A. Moral-Munoz¹

¹ veronica.perezcabezas@uca.es, carmen.ruizmolinero@uca.es, ines.carmona@uca.es, joseantonio.moral@uca.es
Dept. Nursing and Physiotherapy, University of Cádiz, Cádiz (Spain)

² viedma@decsai.ugr.es
Dept. Computer Science and Artificial Intelligence, University of Granada, Granada (Spain)

³ manueljesus.cobo@uca.es
Dept. Computer Science and Engineering, University of Cádiz, Cádiz (Spain)

Abstract

Rheumatology is a wide research area with an extensive background in scientific publications. Thus, the aim of the present study is to identify the highly cited papers in Rheumatology research field, analyzing some aspects such as, the documents distribution by years, journals, authors, institutions, countries and topics. In order to perform the analysis, the H-Classics methodology, based on widely used H-index, has been used. As result it is obtained that Arthritis and Rheumatims is the journal with highest number of documents, with more than half of detected documents. Professor Wolfe, from the University of Kansas, and professor Dougados, from the Paris Descartes University, are the authors with more highly cited papers. The University of California (USA), the University of Boston (USA) and the University of Toronto (Canada) are the main institutional contributors. USA is the main producer, with more than half of the highly cited papers. The present study shows a useful insight into the development and evolution of the Rheumatology research field, revealing actors that have made the biggest research contribution to its development.

Conference Topic
Citation and co-citation analysis

Introduction

Rheumatology is considered as a subspecialty of internal medicine and paediatrics. It includes clinical problems in joints, soft tissues, autoimmune diseases, vasculitis and heritable connective tissue disorders (Cheng and Zhang 2013). In fact, this kind of musculoskeletal diseases are extremely common and have important consequences to the individual and society, being one of the major causes of disease burden around the world (Brooks 2006).

The present study aims at identifying the Highly Cited Papers (HCP) into the Rheumatology scientific production. HCP could be considered important in a research field development because they have attracted the interest of the research community. The concept of citation classic consists in characterize the HCP of a scientific discipline (Garfield 1977). Citations classics help to discover potentially important information towards the development of a discipline, and understand its past, present and future scientific structure.

According to the research literature, a series of papers have been published focusing the bibliometric impact of the Rheumatology research field Cheng and Zhang.
2013) analyzed the articles published in 39 rheumatology journals from 1996 to 2010 using the Scopus database; the number of articles, citations, h-index, and international collaborations were determined by countries or regions. In Chen et al (Chen et al. 2011), the Impact Factors of rheumatology journals from 1999 to 2008 were analyzed and compared with other fields. Cheng and Zhang (Cheng & Zhang 2010) evaluated the scientific production on rheumatology field in the 3 major regions of China (Mainland, Hong Kong, and Taiwan) during the period 2000-2009. Battle-Gualda et al.(Battle-Gualda et al. 1998) analyzed the magnitude, evolution and characteristics of the Spanish scientific production from 1990 to 1996 in Rheumatology. Redondo et al.(Redondo et al. 2016) performed a bibliometric study of the scientific publications on patient-reported outcomes in Rheumatology. However, the HCP of the Rheumatology research field have not been analyzed.

The main aim of the present study is to identify the HCP in the Rheumatology research field using the H-Classic concept. Some aspects can be analysed by identifying the set of highly cited documents: I) the HCP distribution during the period studied; II) the most productive journals, authors, institutions and countries; and, III) the main topics covered by the papers detected.

**Methods**

The set of documents to perform the bibliometric analysis is based on the journals indexed in the Rheumatology category of the Journal Citation Report (JCR-2014), which is used to construct an adequate list of the Rheumatology journals. It has been suggested that JCR contains the most important research documents of the different scientific disciplines, since they are considered as a very important criterion in tenure, promotion and other professional decisions(Hodge and Lacasse 2011; Seipel 2003).

Therefore, in order to develop the HCP analysis, the documents published by the 32 journals indexed in the JCR Rheumatology category were obtained. The search was performed on February 2016. Finally, a total of 96,642 documents (articles and reviews) were retrieved, containing the following information: authors, affiliations, title, year of publication, citations, sources, abstract and keywords.

In order to analyze the HCP, is common to establish a threshold value to discriminate whether a paper is considered as highly cited or not (Garfield 1987; Garfield 1977). In the present study, the concept of H-Classics (Martínez et al. 2014), based on the popular H-index,(Hirsch 2005) is applied to identify the highly cited documents in Rheumatology research field. The concept of H-Classics can be defined as follows (Martínez et al. 2014): “H-Classics of a research area are composed of the H highly cited papers with more than H citations received”.

This approach provides an unbiased and fair criterion to construct a systematic search procedure for highly cited documents. Furthermore, the H-Classics provides a rigorous and scientific method to discover the most relevant papers in a field. Finally, 308 H-Classics (Table 1, full access at [https://doi.org/10.6084/m9.figshare.4648450.v1](https://doi.org/10.6084/m9.figshare.4648450.v1)) were obtained using the approach mentioned above.
Results
In the following sections, the 308 HCP identified (Table 1) in Rheumatology research field are analyzed: I) the distribution of HCP per year of publication is studied, II) journals, authors, institutions and countries producing the highest number of HCP are identified, and III) a content analysis is provided.

Table 1. Highly cited papers (only 10/308 papers are shown, full table can be accessed at https://doi.org/10.6084/m9.figshare.4648450.v1)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Nº Auth.</th>
<th>Paper</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>#4</td>
<td>5</td>
<td>Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation-study of womac - a health-status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug-therapy in patients with osteo-arthritis of the hip or knee. <em>J Rheumatol</em> 1988;15:1833–40.</td>
<td>3131</td>
</tr>
</tbody>
</table>
Distribution of HCP
As afore mentioned, the study was performed with a collection of 96,642 documents, 308 of which were classified as HCP. The distribution of the HCP documents per year is shown in Figure 1. These documents are published between 1957 and 2013, covering a long period of time and concentrating the production in 1990, 1995, and the period 1998-2006. The year with more HCP documents published is 2002. The first document was published by Kellgren et al. (Kellgren and Lawrence 1957) in 1957, and the last document was published by Jennette et al. (Jennette et al. 2013) in 2013. It is worth noting the ranking position of the last document, published in 2013 and ranked 141 (about half of the table). Furthermore, although HCP are usually identified at the beginning of the period, Rheumatology research area is attracting the research community attention in relatively recent years.

![Figure 1. Distribution of Rheumatology H-Classics documents per year of publication.](image)

**Most productive journals, authors, institutions and countries**
Analyzing the social units (journals, authors, institutions and countries) the most productive journals could be detected. In this sense, Table 2 shows the journals with 4 or more H-Classics in the Rheumatology research field. *Arthritis and Rheumatism* with 200 documents is the most productive journal, followed by *Annals of the Rheumatic Diseases* and *Journal of Rheumatology*, with 40 and 26 respectively.

Otherwise, Table 3 shows the authors with 10 or more HCP documents in the Rheumatology research field. Wolfe, F., affiliated with the University of Kansas (USA), and Dougados, M., affiliated with Paris Descartes University (France), are the authors with the highest number of HCP. It is worth noting that 10 authors are from USA, 2 authors from UK, and the rest from France, Canada, Austria, Netherlands and Germany.
Table 2. Most productive highly cited journals.

<table>
<thead>
<tr>
<th>Institution</th>
<th>H-Classes(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arthritis and Rheumatism</td>
<td>205(66.56)</td>
</tr>
<tr>
<td>Annals of the Rheumatic Diseases</td>
<td>40(12.99)</td>
</tr>
<tr>
<td>Journal of Rheumatology</td>
<td>26(8.44)</td>
</tr>
<tr>
<td>Osteoarthritis and Cartilage</td>
<td>8(2.60)</td>
</tr>
<tr>
<td>Arthritis Research</td>
<td>4(1.30)</td>
</tr>
<tr>
<td>British Journal of Rheumatology</td>
<td>4(1.30)</td>
</tr>
<tr>
<td>Rheumatology</td>
<td>4(1.30)</td>
</tr>
<tr>
<td>Seminars in Arthritis and Rheumatism</td>
<td>4(1.30)</td>
</tr>
</tbody>
</table>

Table 3. Most productive highly cited authors.

<table>
<thead>
<tr>
<th>Author</th>
<th>Institution</th>
<th>HCP(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wolfe, F.</td>
<td>University of Kansas, Wichita, Kansas, USA</td>
<td>19(6.17)</td>
</tr>
<tr>
<td>Dougados, M.</td>
<td>Paris Descartes University, Paris, France</td>
<td>19(6.17)</td>
</tr>
<tr>
<td>Felson, D.T.</td>
<td>Boston University, Boston, Massachusetts, USA</td>
<td>16(5.19)</td>
</tr>
<tr>
<td>Emery, P.</td>
<td>University of Leeds, Leeds. UK</td>
<td>16(5.19)</td>
</tr>
<tr>
<td>Hochberg, M</td>
<td>Johns Hopkins University, Baltimore, Maryland, USA</td>
<td>15(4.87)</td>
</tr>
<tr>
<td>Bloch, D.A.</td>
<td>University of Toronto, Toronto, Ontario, Canada</td>
<td>14(4.55)</td>
</tr>
<tr>
<td>Smolen, J.S.</td>
<td>University of Vienna, Vienna, Austria</td>
<td>14(4.55)</td>
</tr>
<tr>
<td>Mchshane, D.J.</td>
<td>Stanford University, Palo Alto, California, USA</td>
<td>12(3.90)</td>
</tr>
<tr>
<td>Breedveld, F.C.</td>
<td>Leiden University, Leiden, Netherlands</td>
<td>12(3.90)</td>
</tr>
<tr>
<td>Hunder, G.G.</td>
<td>Mayo Clinic, Rochester, Minnesota, USA</td>
<td>11(3.57)</td>
</tr>
<tr>
<td>Masi, A.T.</td>
<td>University of Illinois, Peoria, Illinois, USA</td>
<td>11(3.57)</td>
</tr>
<tr>
<td>Burmester, G.R.</td>
<td>Free University and Humboldt University Berlin, Berlin,</td>
<td>11(3.57)</td>
</tr>
<tr>
<td></td>
<td>Germany</td>
<td></td>
</tr>
<tr>
<td>Fries, J.F.</td>
<td>Stanford University, Palo Alto, California, USA</td>
<td>10(3.25)</td>
</tr>
<tr>
<td>Altman, R</td>
<td>University of Miami, Miami, Florida, USA</td>
<td>10(3.25)</td>
</tr>
<tr>
<td>Medsger, T.</td>
<td>University of Pittsburgh, Pittsburgh, Kansas, USA</td>
<td>10(3.25)</td>
</tr>
<tr>
<td>Liang, M.H.</td>
<td>Harvard University, Cambridge, Massachusetts, USA</td>
<td>10(3.25)</td>
</tr>
</tbody>
</table>

On the other hand, according to the information on author addresses contained in the research papers, the corresponding institutions and countries can be identified. Table 4 shows the ranking of institutions with 10 or more HCP. The top 3 most productive institutions are the University of California (USA), the University of Boston (USA) and the University of Toronto (Canada), with 29, 23, and 20 documents, respectively.
Table 4. Most productive highly cited institutions.

<table>
<thead>
<tr>
<th>Institution</th>
<th>Country</th>
<th>HCP(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of California</td>
<td>USA</td>
<td>29(9.42)</td>
</tr>
<tr>
<td>University of Boston</td>
<td>USA</td>
<td>23(7.47)</td>
</tr>
<tr>
<td>University of Toronto</td>
<td>Canada</td>
<td>20(6.49)</td>
</tr>
<tr>
<td>University of Leiden</td>
<td>Netherlands</td>
<td>18(5.84)</td>
</tr>
<tr>
<td>University of Stanford</td>
<td>USA</td>
<td>17(5.52)</td>
</tr>
<tr>
<td>University of Texas</td>
<td>USA</td>
<td>17(5.52)</td>
</tr>
<tr>
<td>University of Leeds</td>
<td>UK</td>
<td>17(5.52)</td>
</tr>
<tr>
<td>University of Harvard</td>
<td>USA</td>
<td>17(5.52)</td>
</tr>
<tr>
<td>University of Pittsburgh</td>
<td>USA</td>
<td>16(5.19)</td>
</tr>
<tr>
<td>Cochin Hospital Paris</td>
<td>France</td>
<td>14(4.55)</td>
</tr>
<tr>
<td>University of Washington</td>
<td>USA</td>
<td>14(4.55)</td>
</tr>
<tr>
<td>University of Manchester</td>
<td>UK</td>
<td>14(4.55)</td>
</tr>
<tr>
<td>Brigham &amp; Womens Hospital</td>
<td>USA</td>
<td>14(4.55)</td>
</tr>
<tr>
<td>University of Vienna</td>
<td>Austria</td>
<td>14(4.55)</td>
</tr>
<tr>
<td>National Institute of Dental and Craniofacial Research</td>
<td>USA</td>
<td>13(4.22)</td>
</tr>
<tr>
<td>University of Kansas</td>
<td>USA</td>
<td>13(4.22)</td>
</tr>
<tr>
<td>University of Erlangen Nurnberg</td>
<td>Germany</td>
<td>13(4.22)</td>
</tr>
<tr>
<td>University College London</td>
<td>UK</td>
<td>11(3.57)</td>
</tr>
<tr>
<td>McMaster University</td>
<td>Canada</td>
<td>10(3.25)</td>
</tr>
<tr>
<td>Mayo Clinic</td>
<td>USA</td>
<td>10(3.25)</td>
</tr>
<tr>
<td>University Nijmegen Hospital</td>
<td>Netherlands</td>
<td>10(3.25)</td>
</tr>
<tr>
<td>University of Michigan</td>
<td>USA</td>
<td>10(3.25)</td>
</tr>
<tr>
<td>University of Alabama</td>
<td>USA</td>
<td>10(3.25)</td>
</tr>
</tbody>
</table>

Finally, Table 5 shows the ranking of countries that produce the HCP in Rheumatology research field. The country with the highest number of HCP is the USA with 156, the half of total number of documents. It is followed by UK and Netherlands with 78 and 56 documents respectively. The predominance of USA in the Rheumatology HCP production is really evident.

It is important to highlight that each document was considered from all authors’ institutions and countries, not only the first or corresponding author.
Table 5. Most productive highly cited countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>HCP(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>156(50.65)</td>
</tr>
<tr>
<td>UK</td>
<td>78(25.32)</td>
</tr>
<tr>
<td>Netherlands</td>
<td>56(18.18)</td>
</tr>
<tr>
<td>Canada</td>
<td>48(15.58)</td>
</tr>
<tr>
<td>Germany</td>
<td>40(12.99)</td>
</tr>
<tr>
<td>France</td>
<td>33(10.71)</td>
</tr>
<tr>
<td>Italy</td>
<td>26(8.44)</td>
</tr>
<tr>
<td>Sweden</td>
<td>23(7.47)</td>
</tr>
<tr>
<td>Switzerland</td>
<td>23(7.47)</td>
</tr>
<tr>
<td>Austria</td>
<td>22(7.14)</td>
</tr>
<tr>
<td>Spain</td>
<td>20(6.49)</td>
</tr>
<tr>
<td>Denmark</td>
<td>14(4.55)</td>
</tr>
<tr>
<td>Belgium</td>
<td>12(3.90)</td>
</tr>
<tr>
<td>Japan</td>
<td>12(3.90)</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>9(2.92)</td>
</tr>
<tr>
<td>Australia</td>
<td>8(2.60)</td>
</tr>
<tr>
<td>Norway</td>
<td>8(2.60)</td>
</tr>
<tr>
<td>Poland</td>
<td>8(2.60)</td>
</tr>
</tbody>
</table>

HCP content analysis

In view of the above information, a conceptual evolution of Rheumatology HCP can be carried out. The first HCP are from the late 50's. Kellgren et al. (Kellgren and Lawrence 1958; Kellgren and Lawrence 1957) wrote the first one on the radiological diagnosis of osteoarthritis, and the second on the frequency of degenerative joint diseases in urban population. Ropes et al. (Ropes et al. 1959) conducted a review of diagnostic criteria in rheumatoid arthritis. From the 60's, only two documents are considered as HCP. One of them (Balazs et al. 1967), based on the parameters of hyaluronic acid in synovial fluid in patients with arthritis, and the second (Mason and Barnes 1969) on the diagnostic criteria of the Behçet's syndrome. From the 70's, several HCP appears with a wide range of topics, such as Lyme's arthritis (Steere et al. 1977), Behçet's syndrome (Shimizu et al. 1979), the American Rheumatism Association classification criteria for gout (Wallace et al. 1977), or a correlational study on four scales for pain measurement (Downie et al. 1978). From the 80's, a total of 44 HCP appear. The HCP first ranked, Arnett et al. (Arnett et al. 1988), appears in this decade and it is about the classification criteria of rheumatoid arthritis. Furthermore, two studies on the rheumatic diseases classification were conducted, one of them by Altman et al. (Altman et al. 1986) centered on knee osteoarthritis, and another by Cassidy et al. (Cassidy et al. 1986) focused on juvenile rheumatoid arthritis. According to our classification, there is great interest in studies centered on different measurement scales and questionnaires about health status, symptoms and disabilities in rheumatic patients. From the 90's, documents are centered in two thematic areas. On the one hand, there are documents focused on the different rheumatic diseases. On the other hand, there are HCP about the questionnaires, indexes and scales validation for measuring the impact and the physical activity on these diseases. In 2002 appears the highest number of HCP. Vitali et al. (Vitali et al. 2002) publishes the document with highest impact, it is about a revised European classification criteria for Sjogren's
syndrome. Furthermore, there are documents focused on the articular cartilage repair (Hunziker 2002; Wakitani et al. 2002), mesenchymal stem cells and their role in rheumatic diseases (Jones et al. 2002), circulatory manifestations and their expression patterns of autoimmune diseases (Cervera et al. 2002; Van Doornum et al. 2002).

From 2002, the number of HCP decreases. The last document appears in 2013, Jennette et al. (Jennette et al. 2013) conducts a review of the International Conference of Chapel Hill (CHCC1994) in the nomenclature of systemic vasculitis, whose goal was to reach a consensus on the names of the most common forms of vasculitis and build a specific definition for each one. It is noteworthy that since 1988 in which appeared a classification of rheumatoid arthritis (rank 1) (Arnett et al. 1988), an update appeared in 2010 (Aletaha et al. 2010), conducted by the American College of Rheumatology in collaboration with the European League Against Rheumatism (EULAR). Also, we should remark that although this document is relatively recent, it appears in a good position (rank 34).

Conclusions

In the present study, Rheumatology HCP have been identified and consequently analyzed using the concept of H-Classics. The analysis of the HCP allows us to highlight the following remarkable findings:

• 308 Rheumatology HCP were identified in the time period 1945-2015, with citation counts ranging from 309 to 9773. The HCP with the highest citations count, about the criteria for the classification of rheumatoid arthritis, is authored by Arnett et al. (Arnett et al. 1988) in 1988.
• Aletaha et al. (Aletaha et al. 2010) in 2010 and Jenette et al. (Jennette et al. 2013) in 2013 (with 907 and 228 cites, respectively), are also remarkable because articles normally need a long period to accumulate citations. The first one is ranked 34 and the second one 141 (about half of the table).
• Arthritis and Rheumatism is the most productive journal with 205 highly cited documents.
• Professor Wolfe, from the University of Kansas, and professor Dougados, from the Paris Descartes University, are the authors with more HCP in the field studied.
• The University of California (USA), the University of Boston (USA) and the University of Toronto (Canada) are the main institutional contributors of HCP.
• The predominance of USA in producing HCP is remarkable. Its production represents more than half of the total of HCP identified.
• The three most cited HCP focus on rheumatoid arthritis, fibromyalgia and osteoarthritis. The research topic that is addressed more within Rheumatology HCP is the diagnostic criteria of different rheumatic diseases, the most frequent are rheumatoid arthritis and osteo-arthritis.

It is worth mentioning the practical application of the present study as it provides a potentially important information to help understand the past, present and future scientific structure of the Rheumatology field that could help its future research development.

Acknowledgments

The authors want to thanks the support of FEDER funds TIN2013-40658-P and TIN2016-75850-R and University of Cádiz project PR2016-067.
References


Retraction: The “Other Face” of Research Collaboration

Li Tang\textsuperscript{1*} Guangyuan Hu\textsuperscript{2} Yang Sui\textsuperscript{3} Cong Cao\textsuperscript{4*}

\textsuperscript{1}litang@fudan.edu.cn
Fudan University, Shanghai (China)

\textsuperscript{2}hu.guangyuan@shufe.edu.cn
Shanghai University of Finance Economics, Shanghai (China)

\textsuperscript{3}suiyanghi@126.com
Kearney A.T., Shanghai (China)

\textsuperscript{4}cong.cao@nottingham.edu.cn
University of Nottingham, Ningbo (China)
East China Normal University, Shanghai (China)

* Authors for correspondence

Abstract
There is an increasing amount of research investigating retractions. Yet little attention has been paid to the relationship between retractions and collaboration. In this paper, we test the hypotheses regarding the relationship between retractions and collaboration on a unique publication dataset of retractions and its control group constructed by the nearest-neighbor-matching approach. Our analysis finds no significant evidence indicating that collaboration suffers from producing flawed research, at least in the form of retraction. We also find that ceteris paribus publications with authors from elite universities are less likely but more quickly to be retracted. There also is no significant impact of collaboration size on the speed of retraction of Chinese articles, although China stands out with the fastest retracting speed. Our findings have policy implications for the governance of global science, especially that involves collaboration.

Conference Topic
research collaboration; retraction; diffusion of responsibility; scientific misconduct
Introduction

Team-dominated knowledge production has become ubiquitous globally. The trend of co-publishing escalating is evidenced by steadily growing team size and proportions of multi-authored publications, many of which are across nation’s borders (LaFollette, 1992; Cronin, 2001; Newman, 2001; Wuchty et al., 2007; Royal Society, 2011; National Academies, 2014). Reasons for producing joint research vary, from greater epistemic authority (Beaver & Rosen, 1979), more easily secured funding (Melin & Persson, 1996), higher-quality work due to cross-pollination of different minds (Cronin, Shaw, & Barre, 2004; Youtie et al., 2013), to possibility of the work receiving more citations (Royal Society, 2011). Meanwhile, the number and the annual rate of retraction of scientific research, or the official declaration of withdrawal of an article from the literature for scientific misconduct or significant errors, also have surged exponentially over the last decade (Steen, 2011a; Van Noorden, 2011; Zhang & Grieneisen, 2012; Riederer, 2014; Sheth & Thaker, 2014). The seeming coincidence, or at least co-concurrence, of the rising collaboration and retraction raises following questions: Is teamwork more likely to be associated with retraction? What factors contribute to the elapsed time between publication and retraction of collaborative work?

Unfortunately, in spite of extant research that examines factors impacting retraction, within our best knowledge, no research has systematically investigated the relationship between retraction and collaboration. This study aims to fill this gap in the literature by examining the effect of collaboration – type and size – on retraction. To do so, we draw on two opposite notions from the social psychological literature on group interaction – diffusion of responsibility and internal auditing – and test our hypotheses regarding the relationship between retraction and collaboration on a unique publication dataset of retractions and its control publication dataset constructed by the nearest-neighbor-matching approach. There is evidence in support of teamwork premium inhibiting retraction; ceteris paribus publications with authors from elite universities are less likely but more quickly to be retracted; and among existing and emerging scientific powers measured by the number of publications, China stands out with the fastest retracting speed. These findings will have policy implications for the enhancement of the governance of knowledge production in collaborative research.

The rest of the paper is structured as follows: Section 2 highlights topic importance and some noteworthy studies on retraction, while Section 3 delineates key notions and hypotheses for testing. Section 4 outlines the data and methodology, followed by analysis in Section 5. Major findings and policy implications are discussed in Section 6.

Revisiting the literature of research retraction

In the knowledge economy, both public policies and individual decisions, ranging from health to education, from parenting to business choices, are now increasingly dependent on the findings from scientific research, thus rendering scientific rigor and research integrity more important than ever (Macfarlane, Zhang, & Pun, 2014). Policy makers and research administrators come to realize that fraudulent findings endanger not only the operation of scientific enterprise but also the wellbeing of the society at large, not to mention wasting rising but still scarce public investment\(^1\) and tarnishing the trust of the public toward science and scientists (Zuckerman, 1988; Lacetera & Zirulia, 2011; Azoulay et al., 2012; Steen 2014). Unsurprisingly, over the last decade, a growing number of studies have investigated the rising phenomenon of retraction, from which four major research streams have emerged.

The first line of research describes the phenomenon. Studies consistently show that both the number and the growth rate of retracted articles have risen sharply over the last decade (Nath et al., 2006), and that such phenomenon has a strong national orientation (Liu & Hu, 2012). While research seems to suggest a high incidence of retractions in articles indexed in PubMed, a biomedical database, anecdotal evidence indicates that publishing misconduct is
significantly greater in non-PubMed articles than in PubMed ones (Zhang & Grieneisen, 2013). Grieneise and Zhang (2012) also argued convincingly that repeat retractors are globally distributed and have skewed distribution at country and institutional levels. Retraction is particularly prevalent for research produced in countries such as the USA, Germany, Japan, China, the UK, and India. For example, China leads in duplicate publications, followed by the USA and India (Grieneise & Zhang, 2012).

The second line of research categorizes reasons for retraction for the sake of policy prevention and intervention. Wager and Williams (2011; 2013) pointed out that reasons for retraction vary, including but not limited to honest but significant research errors, redundant publications, and plagiarism. Steen (2011b) further concluded that about three quarters of the retracted PubMed papers between 2000 and 2010 are due to errors or undisclosed reasons. Fang et al. (2012) refined previous research on causes of retraction. By combining information from Retraction Watch, a blog reporting on retractions of scientific papers, the US Office of Research Integrity (ORI), and other public sources, they reported that only one fifth of the retractions resulted from errors, whereas over two thirds of the papers were removed because of scientific misconduct manifested in falsification, fabrication, and plagiarism. Nath et al. (2006) performed a study to determine how commonly articles had been retracted on the basis of unintentional mistakes and whether such articles had differed from those retracted for scientific misconduct in authorship, funding, type of study, publication, and time to retraction. Their finding is consistent with previous ones, that is, number of retractions was correlated significantly with impact factors of the journal in which the retracted paper was published. Steen (2011b) also argued that journal’s impact factors were often significantly higher for fraudulently produced papers. And Noyori and Richmond (2013) confirmed empirically that the higher the impact factor of a journal, the more cases of retractions there had been.

The third strand of investigation explores the retraction phenomenon from the perspective of scientists. For example, Lacetera and Zirulia (2011) proposed a game-theoretic model to help understand motivations of scientists committing fraud as well as approaches of detecting and preventing fraud. By setting retraction in a dynamic game with asymmetrical information, they found that elite scientists were more likely to commit scientific misconducts but were less likely to be spotted than average misbehaved scientists. They also suggested that more intensive competition might in fact reduce scientific malfeasance because competitors closely monitor new findings.

Recently, studies have appeared treating the event of retraction as an independent variable and examining its impact on the individual and research domain levels. For example, Azoulay et al. (2012) investigated the extent to which “false science” impacts the rate and the direction of scientific change. Lu et al. (2013) also looked at how the retraction event impacted citations through comparing an expanded treatment group, which includes not only retracted articles but also prior articles of the retracted authors, and a control group, which consists of papers with similar citation patterns to treated papers prior to the date of retraction. They found that citation penalty for non-self-reporting retraction goes beyond the retracted paper itself. Along this line to investigate the impact of retraction on the citations to retracted authors’ prior work, Jin et al. (2013) documented a heterogeneous impact of retraction penalties on eminent and less-famous collaborators. The reverse Matthew Effect, in their words, hints at protection of established reputations.

The current study continues these inquiries following especially research streams one and three. One research gap that we spotted is a possible connection between co-publishing and retraction. In a departure from past scholarship, we try to understand the mechanism that governs the behavior of individuals in research collaboration. In particular, we are interested in knowing whether working together helps encourage responsible publishing effort.
Theories and hypotheses

As an old proverb goes, too many cooks spoil the broth. In the social psychology literature, diffusion of responsibility, also referred to as bystander effect, suggests that an individual is less likely to take responsibility for action or more likely to be idle with the presence of others, as the individual assumes that others either are responsible for taking action or have already done so (Darley & Latane, 1968; Forsyth et al., 2002). This bystander inaction could also occur in real-life academia. In the case of joint publication, for example, it is reasonable to assume that each coauthor feels that the responsibility for credibility and quality of the coauthored work is diffused, thus not necessarily taking care of the validity of the collective knowledge product.

A closely related but different notion is social loafing (Williams & Karau, 1991). That is, team members can become disgruntled and de-incentivized by unfair workload distribution due to social loafing (Tsai & Chi 2008). In the case of scientific collaboration, social loafers contribute less than their fair share to collective efforts but reap the benefits of the efforts of group members as the entire group is rewarded or punished by new knowledge demonstrated in the joint publication (Aggarwal & O’Brien, 2008; Smalheiser et al., 2005).

This underperformance due to diffusion of responsibility and social loafing, combined with the cost of collaboration (such as knowledge fragmentation and coordination failure), thus increases the likelihood of producing flawed scientific findings, which, in return, leads to a higher probability of retraction. Accordingly, our first hypothesis is as follows:

- **Hypothesis 1a**: As the size of the coauthorship increases, there will be greater incidence of retraction.

In contrast to diffusion of responsibility embedded in the social loafing theory, internal auditing, which is linked to the social interdependence theory (Johnson, 2003), provides an alternative scenario. Conceivably, due to internal auditing, more coauthors mean a higher likelihood of carrying out more scrupulous checking and knowledge validation and ensuring higher standard of quality control, thus leading to robust research that is less likely to be retracted. In other words, without collaborators’ internal auditing, a single author might be more likely to produce sloppy or even false work. So, our alternative hypothesis is:

- **Hypothesis 1b**: The size of coauthorship is negatively associated with the possibility of retraction holding other factors constant.

Factors impacting retractions

While Pozzi and David (2007) reported a lag of three years for investigated retractions by the US ORI, Redman et al. (2008) documented a much shorter retraction time – 20.75 months on average – in their study of the 315 retracted papers in the 1995–2004 PubMed data. However, existing research has paid very little attention to the factors impacting the time between publication and retraction, with few notable exceptions. A recent study conducted by Furman and colleagues (2012), for example, argued that no observable factors impacted time to retraction except for publication year, whose statistically significant and negative regression coefficient provided strong evidence in support of the trend of shortened times of detecting flawed findings.

Intuitively, multi-authored research on average receives more scholarly attention and scrutiny and thus is possibly quicker to be detected for its shaky or fraudulently produced findings. So our second null and alternate hypotheses are:

- **Hypothesis 2a**: The size of coauthorship is negatively associated with the elapsed time between publication and retraction.
- **Hypothesis 2b**: The size of coauthorship is positively associated with the time between publication and retraction.
Methodology
In order to test our hypotheses, we first constructed datasets in a series of sequential steps. Our primary source is a retracted paper dataset retrieved from Thomson Reuters’ Web of Science (WoS), an index of 11,600 peer-reviewed journals world-wide with coverage spanning a wide range of scientific disciplines. To develop this dataset, we started with using a composite Boolean query to search retraction notices, from which we eventually identified 2,087 unique retracted papers then downloaded the full bibliographical records of the retracted articles indexed in WoS from 1978 to 2013. Only original research articles were included in our dataset. We consider this a more up-to-date dataset than not only that built on PubMed, the database adopted by most previous studies, but also that utilizing information from WoS. Therefore, our dataset reveals a more comprehensive picture of the retraction phenomenon and especially the factors that have impacted retractions.

Based on the nearest-neighbor-marching principle proposed by Furman et al. (2012), we additionally identified two control articles for each retracted one by choosing its nearest neighbors. We started with the two articles immediately before and after the retracted article in the same issue of the same journal in which the retracted article was published. If neither one is qualified (for example, its document type is conference abstract, letter, correction, and editorial, among others), we next tried its nearest neighbor. The farthest neighborhood distance is 3, i.e. three papers ahead of or behind the retracted one, and we stopped search if we still were unable to locate a matching article. If a retracted article was the first or last one in that issue, we only included one comparison. In this way and with several rounds of data cleaning and standardization, we finally identified 3,970 control records with 96.6% matching rate. This approach of matching baseline of journal and publication time has an explicit and effective merit as it holds constant other factors that might have affected the incidence of retraction.

The two datasets were then imported into the text mining software VantagePoint. Our final core dataset for analysis consists of 6,057 records with 2,087 retracted articles and 3,970 associated control matched articles. We further retrieved journal impact factors from the 2013 ISI Journal Citation Reports (JCR) and global rankings of the institutional affiliations of the authors of the retracted and control articles from the 2014 Academic Ranking of World Universities (ARWU) released by Shanghai Jiaotong University and merged them into our dataset.

Analysis

Descriptive analysis
Our data indicate that between 1978 and 2013, both retraction quantity and rate increased with time. Measured by the year when the retracted articles were published, only three retractions appeared in 1978 publications, but retractions rose to 198 in 2000 and 213 in 2010. According to the year of retraction notices, the frequency of retraction appears to be precipitous, yet the publications between 2010 and 2013 were retracted more than ten times those between 2000 and 2003.

We calculated the subject-specific quartile impact factors of each journal based on the 2013 ISI JCR. If a journal is assigned to different subject categories or disciplines, we took its impact factors at both the highest and lowest quartiles (Liu and Hu, 2016). Consistent with previous findings, noted above, our analysis indicates that retractions appeared substantially more frequent in journals with higher impact factors. As illustrated in Figure 1, even in the pessimistic mode of taking the lowest quartile, 46% of the retracted articles were published in Quartile-1 (high-impact-factor) journals while only 11% in Quartile-4 (low-impact-factor).
journals. In the optimistic mode of allocating journal’s impact-factor quartiles, the retractions were 55% in Q1 but 7% in Q4.

Figure 1 Distribution of Journal Impact Factors of Retracted Articles by Quartile

With regard to research domains, retraction was more common in hard sciences, especially in the biomedical and life sciences, similar to Lu’s findings (2013), although Lu’s research only focused on post-2000 WoS retracted articles. As illustrated in the inner circle of Figure 2, over 60% of the retracted articles were in the life sciences & biomedicine; by sharp contrast, only 0.1% of the arts & humanities papers and 5.1% of the social sciences papers were retracted. The highly uneven distribution of retractions across disciplines may reflect possible lower incidences of false science or lower rates of detection of problematic research in the arts & humanities and social sciences where knowledge validation norms may differ. We then benchmarked the distribution of retracted articles against all publications in five research areas between 1990 and 2013 (shown in the outer circle of Figure 2). Apparently, the proportion of retracted life sciences & biomedicine articles is one third more than their share in the WoS articles, which only contributed 42.5% of the indexed publications. The leading scientific nations of the USA, China, Japan, Germany, and India all witnessed a higher proportion of retractions in the life sciences & biomedicine relative to their shares of papers in this research domain. The enormous consequences and economic potential of the research in these fields as well as fierce competition for positions, promotions, funding, and especially priority of discovery and peer recognition might lead life scientists to rush to publish, thus also subjecting their research to stricter scrutiny.

The distribution of flawed research also is highly skewed nationally. The top five countries in terms of the number of retracted articles were the USA (622), China (341), Japan (263), Germany (184), and India (141). As shown in Figure 3, the USA published and
retracted almost similar percentages of papers; China, Japan, and Germany each produced some 7% of the global WoS papers, but China and Japan retracted much higher percentages of papers than Germany; and the India’s share of its retracted papers was 2.5 times its share of the global WoS papers.

Table 1 compares the collaboration size (number of authors, number of affiliations, and number of countries) and time to retraction for retracted vs. control paper groups. Apparently, at all three dimensions of author, affiliation, and country, retracted articles have smaller scopes of collaboration. We particularly took a close look at China by extracting out a sub-sample of 341 retracted articles matched with 629 comparisons. As shown in Figure 4, there is no significant difference in collaboration size between the Chinese and global retractions, but the time to spot and retract flawed research involving Chinese researchers was
Figure 3. Article distribution by countries: Retractions vs All WoS articles

Note: Records retrieved and calculated by authors. The number and proportion of WoS articles in countries were retrieved and analyzed online on June 18, 2016.

Table 1 Collaboration size and time to retraction between retracted and control groups

<table>
<thead>
<tr>
<th></th>
<th>Number of Authors</th>
<th></th>
<th>Number of Affiliations</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Mean</td>
<td>Median</td>
<td>Max</td>
</tr>
<tr>
<td>Retracted</td>
<td>1</td>
<td>4.95</td>
<td>4</td>
<td>32</td>
</tr>
<tr>
<td>Control</td>
<td>1</td>
<td>5.36</td>
<td>5</td>
<td>53</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Number of Countries</th>
<th></th>
<th>Time to Retraction (months)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Mean</td>
<td>Median</td>
<td>Max</td>
</tr>
<tr>
<td>Retracted</td>
<td>1</td>
<td>1.24</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Control</td>
<td>1</td>
<td>1.29</td>
<td>1</td>
<td>12</td>
</tr>
</tbody>
</table>
much shorter: the median time for the Chinese retractions was 12 months, which is about half of that for the global retractions. Next we examined whether such difference is statistically significant controlling for other factors.

Regression analysis

Dependent variables

Our unit of analysis is article. The two key focal variables of the study are 1) retraction of flawed research, and 2) retraction time lag. The first dependent variable, retraction, is a dichotomous variable: an article is coded 1 if it is in the retraction group and 0 if in the control group. The second dependent variable is time to retraction, a continuous variable measured by the natural log of the elapsed months between an article’s publication and retraction\textsuperscript{11}.

Explanatory variable

The major independent variable of the study is collaboration size. Following common practices, we measured collaboration size by the following three indicators:

• Number of authors: numerical variable, number of authors
• Number of affiliations: numerical variable, number of unique affiliations
• Number of countries: numerical variable, number of unique countries

Control variables

In addition to year of publication, research domain, and journal’s impact factor, which were explicitly controlled by our nearest-neighbor-matching approach, as explained, we also controlled for research environment by adding a dummy variable indicating affiliation of any primary author (i.e. first author and reprint author) or any author with a global elite or Top-100 university on the 2014 Academic Ranking of World Universities. The logic is very simple: scholars from elite universities are more likely to care about their academic
reputations and those of their institutions as well as the devastating ramifications of retraction for their career. Additionally, we added a set of primary country dummy variables to the regression models to control for the research culture factor.

For testing Hypothesis 1 we used logistic regressions while OLS regressions using the natural logarithm of the time-to-retraction in months were adopted for testing Hypothesis 2. Our primary results are presented in Tables 2 and 3 (in both tables, Panels 1 and 3 are for the global dataset while Panels 2 and 4 zoom in the China sub-dataset). Whole counting was adopted here, as there is no one-to-one link between authors and their reported institutions for multi-authored publications indexed in WoS (Agrawal, McHale, & Oettl, 2013).

Table 2 Factors affecting the likelihood of retractions by logistic regressions: 1978–2013

<table>
<thead>
<tr>
<th>Panel 1: Odds Ratio [95% Conf. Interval]</th>
<th>Panel 2: Odds Ratio [95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Authors</td>
<td>0.95[0.94-0.97]***</td>
</tr>
<tr>
<td>Number of Affiliations</td>
<td></td>
</tr>
<tr>
<td>Number of Countries</td>
<td>0.77[0.67-0.88]***</td>
</tr>
<tr>
<td>Author Top-100 Univ.</td>
<td></td>
</tr>
<tr>
<td>PrimaryAuthor USA</td>
<td>1.12[0.97-1.30]</td>
</tr>
<tr>
<td>PrimaryAuthor China</td>
<td>2.19[1.82-2.62] ***</td>
</tr>
<tr>
<td>PrimaryAuthor Germany</td>
<td>1.66[1.32-2.10] ***</td>
</tr>
<tr>
<td>Control for Publication Year</td>
<td>Yes</td>
</tr>
<tr>
<td>Control for Research Area</td>
<td>Yes</td>
</tr>
<tr>
<td>Control for Journal Impact Factor</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>6054</td>
</tr>
<tr>
<td>Pro&gt;chi²</td>
<td>0.0000</td>
</tr>
<tr>
<td>Pseudo R²</td>
<td>0.03</td>
</tr>
</tbody>
</table>

**Note:** * Significance at 10% level; ** Significance at 5% level; *** Significance at 1% level.

641
Table 3 Factors affecting the time-to-retraction by using OLS regressions: 1978–2013

<table>
<thead>
<tr>
<th>Panel 3: Coef. (Std. Err.)</th>
<th>Panel 4: Coef. (Std. Err.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global-Model 4</td>
<td>Global-Model 5</td>
</tr>
<tr>
<td>Number of Authors</td>
<td>0.02 (0.01)**</td>
</tr>
<tr>
<td>Number of Affiliations</td>
<td>0.03 (0.02)*</td>
</tr>
<tr>
<td>Number of Countries</td>
<td>0.04 (0.04)</td>
</tr>
<tr>
<td>Author Top-100 Univ.</td>
<td>-0.13 (0.05)**</td>
</tr>
<tr>
<td>PrimaryAuthor USA</td>
<td>0.05 (0.06)</td>
</tr>
<tr>
<td>PrimaryAuthor China</td>
<td>-0.11(0.06)*</td>
</tr>
<tr>
<td>PrimaryAuthor Japan</td>
<td>0.64(0.07)***</td>
</tr>
<tr>
<td>PrimaryAuthor Germany</td>
<td>0.36(0.08)***</td>
</tr>
<tr>
<td>PrimaryAuthor India</td>
<td>0.17(0.08)**</td>
</tr>
<tr>
<td>Control for Year</td>
<td>Yes</td>
</tr>
<tr>
<td>Control for Research Area</td>
<td>Yes</td>
</tr>
<tr>
<td>Control for Journal Impact Factor</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>2086</td>
</tr>
<tr>
<td>Pro&gt;F</td>
<td>0.0000</td>
</tr>
<tr>
<td>Adj. R²</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Notes: * Significance at 10% level; ** Significance at 5% level; *** Significance at 1% level.
For robustness testing, we also test if the first author or the corresponding author is affiliated with any of the Top-100 or Top-50 research universities. The results are similar.

Three findings on global retractions are worthy of noting (see Panel 1of Table 2 and Panel 3 of Table 3). To begin with, Hypothesis 1a is not supported. Instead, we found suggestive evidence in support of internal auditing occurring. The odds ratios of collaboration size are less than 1 and statistically significant at authors and affiliations, indicating that collaboration size is negatively associated with the retraction event holding other factors constant. All else being equal, one additional coauthor increases the odds of being retracted (versus not being retracted) by a factor of 0.95; one additional collaborated institute increases the odds of being retracted by a factor of 0.92. In other words, given that the diffusion of responsibility is more likely to occur under conditions of anonymity, we are happy to report that collaborative research does not necessarily mean to be dysfunctional. The logic is simple that a larger collaboration size means wider channels of knowledge diffusion and validation and therefore a higher probability of being identified for fraudulent findings or significant errors. Or, there is internal auditing at work. Hypothesis 2a is not supported either. As shown in Panel 3 of Table 3, collaboration size measured by number of authors and number of affiliations has statistically positive impact on retraction time, but such impact is insignificant for number of nations (see Global-Models 4–6).

Second, our data suggest that publications with authors from elite universities were less likely to be retracted. In the global dataset, only slightly over one in five (22.6% to be exact) retracted articles involved at least one author from a Top-100 university, and 14.9% of the
retracted articles had a first or reprint author from a global elite university. As shown in Global-Model 1 of Table 2, the odds-ratio for authors affiliated with a Top-100 university to withdraw a paper is 0.77. That is, holding other factors constant the odds of an article with at least one author from a Top-100 university being retracted were 23% lower compared with one without such an author, which suggests a more pronounced premium of global elite universities in inhibiting flawed science through collaboration. The same pattern also holds for collaboration size measured by number of institutional affiliations and number of countries. Meanwhile, if a paper enlisting an elite scientist had to be retracted, it was retracted quicker (see Global-Models 4–6 of Table 3).

One speculation for this phenomenon is that collaboration involving leading scientists is likely to produce findings at the frontier so as to be scrutinized by a larger community for their validity. Therefore, flaw, if any, is also quicker to be detected. This result is different from those of Furman et al.’s (2012), which, controlling for first-year citations and affiliation of authors with a Top-25 US university, concluded that neither retraction nor its quickness is correlated with the characteristics of the article, article’s author, or author’s institution. Furthermore, researchers, especially those at renowned universities, care more about their reputations, thus often assuming rather than diffusing the responsibility of validating the collective product.

The third and final interesting finding concerns collaboration involving scientists from the countries of interests. As demonstrated in Panels 2 and 4, among the top five countries with the largest number of retractions, ceteris paribus, China and India stand out with the largest likelihoods of retraction of the research involving their scientists as primary contributors, and China also is the very country with the fastest retracting speed. Adjusting for contributions made by the other variables in the model, the likelihoods of articles with a primary author from China and India being retracted were 1.96 and 2.65 times those of articles with a primary author from the US respectively. This finding supports Segal’s statement (2011) that the emerging scientific powers whose goals of intimately linking scientific research to the national pride of catching up with and surpassing the incumbents sometimes have ended up with unintended and mostly undesirable consequences.

Panel 3 suggests that among the top five countries with the largest number of retracted articles, only articles with primary authors from China demonstrates negative association with time-to-retraction. We further zoomed in the Chinese case, which consists of 341 retractions and 629 corresponding matches. Opposite to the findings from the global dataset, China’s internationally collaborated articles are more likely to be retracted compared to the work produced solely by its domestic researchers (China-Model 3 of Table 2). This does not suggest that collaborative research between domestic Chinese scientists is less problematic but only means that the rate of spotted retractions is less for such research than the international collaboration involving scientists from China. And the inhibiting role of elite authors on flawed research is even more substantially evident for Chinese retractions: the odds of an article with at least one author from a Top-100 university being retracted were 52% (i.e. 1–0.48) lower than the work involving no such an author. We also did not observe significant impact of collaboration size on the speed of retraction of Chinese articles. But as shown in China-Models 4–6 of Table 3, again, once China’s flawed research was spotted, the involvement of authors from a global elite university speeded up retraction significantly.

Conclusion and Discussions
Drawing upon two different social psychological notions – diffusion of responsibility and internal auditing – on team effectiveness and applying them to retractions of collaborative research, we provide empirical evidence on the relationship between collaboration and
retraction and examine the driving factors and the speed of retraction. Our findings especially contribute to the ongoing debate on factors impacting retraction at two fronts. First, given its importance as a driver for publication quantity and visibility, collaboration is worth studying in relation to retraction. In fact, to our knowledge, this is the first study coming up with evidence in support of the premium of team effect on deterring flawed research. Second, collaboration, especially involving elite scientists, seems to be more effective in helping spot then discourage, if not prevent, flawed research. Nevertheless, rising retractions, including those from collaboration involving scientists from elite institutions, also point to the existence of social loafing. There would have been less retractions had the flaw been detected in the stages of research, preparation of manuscripts for submission or internal checking.

As well as inspired by the prior scholarship (Azoulay et al., 2012; Furman et al., 2012; Lu et al., 2012; Jin et al., 2013), our research brings a new perspective – collaboration – into the literature on research retraction. We adopted the same nearest-neighbor-matching approach (Furman et al., 2012) and tried to analyze the same two factors that correlate with the retraction event and the retraction time lag. While Furman et al. (2012) focused on medical sciences using data from PubMed between 1977 and 2006, our analysis covered broader fields with data retrieved from WoS for a longer period of time between 1978 and 2013, thus extending the study of the retraction phenomenon to a much larger retracted group (2087 vs. 677). While Furman et al. (2012) controlled for Top-25 US universities, among others, we examined the global retraction phenomenon by controlling for the involvement of authors from global Top-100 universities, and explicitly measured collaboration through three indicators: number of authors, number of affiliations, and number of countries. As such, in contrary to one of their major findings that retraction rate is “uncorrelated with article, author, or institution characteristics,” we found that publications with authors from elite universities are less likely but quickly to be retracted.

Indeed, in addition to collaboration, other factors also may be associated with the likelihood and speed of retraction. But our methodological innovation – both the nearest-neighbor-matching approach and the use of logistic regression and OLS – allows us to control for such factors as research domain, publication year, journal visibility, research environment and culture. Although we are not seeing “direct” linkage between collaboration and retraction, our findings provide suggestive evidence to lend support that working together helps encourage responsible research behavior among researchers. As such, we are able to confidently conclude that even controlling for other compounding factors, on average, collaborated research is less likely to be “false science”; and that such “false science,” once so labeled, is quicker to be retracted. Of course, the number/type of collaboration itself could be endogenous and driven by the factors mentioned above as well as other factors such as easy access to facilities/resources, validation of knowledge, inherently interdisciplinary and exceptionally explorative nature of the research at the frontier, or increased efficiency of knowledge production due to division of labor (Beaver and Rosen, 1979; Uzzi & Spiro, 2005). Given that this empirical study is based upon secondary data, we can only speculate explanations within our knowledge of research governance practice and extant literature.

As our interest in the relationship between collaboration and retraction is partly driven by the co-concurrence of increasing dominance of team science and increasing invalided knowledge, our findings have policy implications for the governance of global science, especially that involves collaboration. To begin with, scientific misconduct is becoming a major horny problem that has bedeviled the research community for the last couple of decades. Therefore, globally, there is a compelling government interest in promoting responsible research behavior; and researchers taking responsibility in collaboration is one of the ways to inhibit false science. Our study suggests that jointly published research, especially with contributions of primary authors from top universities, is less likely to be retracted. This finding offers

644
empirical support for policy proposals that endorse research collaboration, especially that with elite scientists at top universities.

Second and related, our findings are particularly relevant to current performance evaluation policy in some countries that highly de-incentivizes collaboration (Zhou, Thijs, & Glanzel, 2009; Yan, Rousseau, & Huang, 2015). Our analysis suggests that collaboration is positively correlated with the probability of reducing flawed research. China is now the second largest knowledge producer. Yet, in many Chinese universities only the status of first author or reprint author is counted toward faculty tenure and promotion (Tang, Shapira, & Youtie, 2015). For example, the written criteria for tenure and promotion at some institutions, such as Shanghai Jiaotong University, clearly state that to be eligible for promotion to a higher academic rank a faculty member must publish two first- or reprint-authored papers in WoS-indexed journals. This means that any paper in which an academic is listed as a second or other author will be highly discounted if not completely excluded. Therefore, academics are understandably reluctant to collaborate if they are not listed as the primary authors, especially with competitors within the same college. This has suffocated internal collaboration (Wang et al., 2015). If, however, Chinese universities change policies to give credits to all authorship toward tenure or to count all authorship reasonably, this would incentivize academics to collaborate. This may also invite free ridership. Therefore, some universities, such as the Shanghai University of Finance and Economics, have adopted sophisticated fractional counting formulas for credit sharing to reduce the existence of possible ghost authors. Of course, collaborators need to understand that the privilege of authorship comes with not only credits but also responsibility.

Third, our empirical analysis demonstrates that collaborative publications with Chinese or Indian scientists as primary authors are far more likely to be retracted than those led by scientists from other major countries. This is consistent with the general concern about developing and emerging scientific countries such as China and India that still lack intellectual capital but strive to seek their seats at the league table of global academia (Zhou & Leydesdorff, 2007; The Economist, 2012; Liu, et al., 2013; Liu, et al., 2015; Tang & Hu, 2013). Therefore, the Chinese and Indian governments should have incentives to tackle the frequent and rising incidence of problematic research involving their scientists as such research has severely damaged the reputation of not only their scientists but also their countries. Scientists from these countries need to make extra efforts to maintain the integrity of their work as problematic research, if not stopped, will only discourage collaboration and delay the process through which these countries pursue scientific excellence.

Fourth, notwithstanding unsupported evidence of diffusion of responsibility in publishing retractions, we appeal for authors and journal articles to explicitly state who has contributed what in collaborative research.

Before concluding, a few reminders on the caveats seem necessary. Retracted articles do not necessarily entail scientific misconduct; unintentional but significant errors can also lead to retraction. In this paper, we have not examined the reasons for paper retraction. It is conceivably fine for this particular work as, regardless of its heterogeneous causes, retracted articles are flawed research, or “false science,” a shaky/unsteady stepping stone for later follow-on investigations (Mokyr 2002; Azouley et al., 2012).

Acknowledgments
This research draws on support from the National Natural Science Foundation of China (#71303147) and Shanghai Soft Science Program (#14692102900). Cong Cao’s work was supported by East China Normal University’s High-Level Overseas Expert Scholarship. The conclusions contained herein are those of the authors and do not reflect the views of funders.
References


---

1 For example, Steen (2014) established that on average each retracted article funded by the National Institutes of Health between 1992 and 2012 incurred an average direct cost of $392,582.
2 The ranking order of these countries changes a bit depending on the coverage of retraction articles analyzed.
3 Different samples and coding schemes are the two main reasons accounting for the different results between Steen (2011b) and Fang et al. (2012). First, while both used PubMed as data sources, Steen (2011b) covered about 800 retracted articles between 2000 and 2010 and Fang et al. (2012) looked at some 2000 retracted articles between 1973 to May 2012. Second, Steen (2011b) collapsed “undisclosed reasons” of retraction to the category of errors based on retraction notices, while Fang et al. (2012) improved the research by combining information from ORI. Please note Steen also is a co-author of Fang et al. (2012).
4 We searched “retract*” in the fields of title, key words and abstract and confined to the document type of corrections for the period from 1978 to 2013. For each retrieved hit, we linked to its retracted article. After several rounds of independent verification and cross-checking of two team members, 2,087 unique retraction notices and their corresponding retracted articles were identified and downloaded in January 2014.
5 Among them 60 retracted articles have one single match for each.
6 Considering the journal coverage dynamics of WoS, if the impact factor is not available in JCR 2013, we used its latest available impact factor.
7 For consistency, non-university institutions such as the Chinese Academy of Sciences and US national labs were treated as non-elite.
8 For example, Acta Neurochirurgica (ISSN: 0001-6268) had an impact factor of 1.766 in 2013, which ranked it the 124th among 192 journals in the field of clinical neurology, i.e. the third quartile in the pessimistic mode, and 84/198 in surgery, i.e. the second quartile in the optimistic mode.
9 Only four document types – article, review, note, and letter – are considered in our record retrieving and calculation.
10 Among 341 retractions involving scholars from China, 310 have Chinese primary authors, including 303 reprint authors and 7 first authors.
11 There are 108 retracted articles with missing values of the publication month for either retracted article or retracted notice. We estimated their elapsed months by assuming the same month of publication and retraction.
12 Furman et al. (2012) confined elite universities to Top-25 or Top-50 US ones.
13 i.e. 2.19/1.12=1.96; and 2.97/1.12=2.65.
For robustness tests, we set a stringent criterion of any primary author from a Top-100 university involved in retraction (Please see Appendices 2 and 3), and similar patterns hold. Additionally, we also conducted the following robustness tests: Recode three collaboration size variables into dummy variables, i.e., single-authored paper vs. multi-authored paper, intra-organization collaboration vs. inter-organization collaboration; domestic paper vs. internationally collaborated paper; and added the squared terms of three collaboration size variables into the model. Such tests did not change the major findings. Results are available upon request.

The concern is comparatively less as Chinese researchers do not compete against international partners for promotion and research funding. But still secondary authorship is highly undervalued for Chinese researchers in the situation of international collaboration.
Abstract
There are three main reasons for retraction: (1 ethical misconduct (e.g. duplicate publication, plagiarism, missing credit, no IRB, ownership issues, authorship issues, interference in the review process, citation manipulation); (2 scientific distortion (e.g. data manipulation, fraudulent data, unsupported conclusions, questionable data validity, non-replicability, data errors – even if unintended); 3) administrative error (e.g. article published in wrong issue, not the final version published, publisher errors). The first category, although highly deplorable has no effect on the advancement of science, the third category is relatively minor. The papers belonging to the second category are most troublesome from the scientific point of view, as they are misleading and have serious negative implications not only on science but also on society. In this paper, we explore some temporal characteristics of retracted articles, including time of publication, years to retract, growth of post retraction citations over time and social media attention by the three major categories. The data set comprises 995 retracted articles retrieved in October 2014 from Elsevier’s ScienceDirect.

Conference Topic
Scientific fraud and dishonesty; citation analysis; Altmetrics

Introduction
Retracted articles, although only a tiny fraction of the total number of articles published raise considerable interest and can have both scientific and societal implications (e.g. the article hinting that MMR vaccinations can cause autism – Wakefield et al., 1998 – not included in the reference list on purpose, which was retracted after 12 years and caused parents to refuse to vaccinate their children). Retractions can also affect the lives of the authors of retracted articles – dismissal from workplace, and in the most extreme case even committing suicide (see Cyranoski, 2014).

There are a large number of studies on retracted articles examining different aspects of the phenomenon (e.g Cokol, Ozbay, & Rodriguez-Esteban, 2008; Steen, 2011; Fang, Steen & Casadevall, 2012; Fanelli, 2013; Budd, Coble & Abritis, 2016; Bar-Ilan & Halevi, 2017). Several studies suggested to categorize reasons for retraction (e.g. issues with data quality and/or interpretation of data; ethical misconduct; data fabrication or falsification; plagiarism; duplicate publication; publisher error; authorship disputes and copyright infringement (Grieneisen & Zhang, 2012; Nath, Marcus & Druss, 2006; Singh et al., 2014; Wager & Williams, 2011).

Previous studies alerted on continued citations of articles after retraction. We call these “post-retraction citations”. The dataset used for this study is comprised of 995 articles that were retracted on or before October 2014. The number of citations these articles received was recorded in October 2014, January 2016 and April 2017, allowing us to examine the temporal trends in post retraction citation. In addition, Mendeley reader counts were also collected at these dates, which serves as an altmetric indicator of interest in scientific publications by a
wider community of people interested in science, but are not necessarily authors of scientific publications (Bar-Ilan, Shema & Thelwall, 2014; Haustein et al., 2014). The reasons for retraction were classified into a number of specific categories along the lines of previous studies mentioned above. The specific categories were collated into three major categories: (1) Ethical misconduct (2) Scientific distortion and (3) Administrative error. The temporal trends were analysed per major categories, highlighting the differences between the categories.

Data Collection
ScienceDirect, Elsevier’s full text database was accessed in October 2014. The database was queried for the term “RETRACTED” in the article title. In ScienceDirect, each retracted article is preceded with the word “RETRACTED”. A total of 1,203 results retrieved from which 995 were retracted articles. The results that were retraction notices, duplicates and papers whose original titles included the word "retracted" were excluded. Each retracted article is accompanied with a retraction notice which explains who retracted article and the reason for retraction. Usually, but not always, the reason for retraction – called Retraction Notice – is published in the same journal at or near the time of retraction. From the year of publication of the retraction notice the year of retraction was deduced. If no separate retraction notice was published, the year of retraction could not be established (this happened for 218 papers (21.90%) in our dataset.

A classification of reasons for retraction, based on earlier studies was developed. The content of retraction notices was analysed by both authors and classified into 18 specific categories independently (there was 92.5% agreement between the two coders). The specific categories were collapsed into three major categories with a 96.28% agreement between the coders:

Ethical misconduct which includes:
- Authorship disputes, citation manipulation, copyright/legal issues, duplicate publication, plagiarism, missing credit, review fabrication, unauthorized data reuse and other ethical issues (e.g. no IRB approval)

Scientific distortion which includes
- Data errors (intentional or unintentional), data fabrication, data integrity, data manipulation, data cannot be validated, findings not replicable, wrong interpretation of results

Administrative error which includes
- Final version of the article was published, wrong issue, etc.

The reason for differentiating between ethical and scientific issues is that although ethical misconduct is deplorable and is contradictory to the norms of science (Merton, 1942), it has no or only minor effect on the advancement of science. The usual name for the second category is scientific misconduct, however we preferred the term “scientific distortion”, as this category includes also data errors which might be unintentional. In either case, whether intentional or unintentional, articles belonging to this category are hurdles for the advancement of science, as they mislead scientists who rely on the results of such articles. The third category is a minor category covering other reasons for retraction that have no influence on the advancement of science.
Results and Discussion

Years to retract

It is well known that retractions take time as can already be seen in Figure 1, where the earliest retracted article in our dataset is from 1985, while the earliest retraction notice is from 1998 (note that the retraction date of 218 retraction notices could not be established). Sometimes problems with the article are not recognized immediately and are disputed, other times committees are set up to decide on the fate of the article (or articles if the investigation is against a person and not a specific paper), if a paper has multiple authors it takes time to agree to request a retraction and there can also be legal complications. COPE – the Committee on Publishing Ethics (Elsevier is a member of COPE) has clear guidelines on retractions (Wager, Barbour, Yentis & Kleinert, 2009). RetractionWatch (http://retractionwatch.com/) and PubPeer (https://pubpeer.com/, see also https://en.wikipedia.org/wiki/PubPeer) are “watchdogs” of scientific distortion and ethical misconduct.

Figure 1. Publications vs. retraction over time

Table 1 provides information on the number of years to retract an article by category for the articles for which retraction date could be established. More than 50% of the articles are retracted within one year, but there are differences by categories, scientific misconduct takes longer and administrative errors are corrected within 5 years or less. Note that there is a single article that was retracted after 28 (!) years form 1985 till 2013, the retraction was requested by the authors on ground of “duplication of data publication and text re-cycling “. There were twelve articles with retraction date after 2014, when they were already retrieved as “retracted”, indicating that it takes time to publish a separate retraction notice.

In total, there are 995 articles in the data set, 218 articles without retraction date, and when these are also included then the distribution between the categories becomes:

- 642 articles retracted because of ethical misconduct (65%)
- 302 articles retracted because of scientific distortion (30%)
- 51 articles retracted because of administrative reasons (5%)
Table 1: Years to retract - total and by categories

<table>
<thead>
<tr>
<th># years</th>
<th>All Categories</th>
<th>Ethical Misconduct</th>
<th>Scientific Distortion</th>
<th>Administrative Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>154</td>
<td>104</td>
<td>37</td>
<td>13</td>
</tr>
<tr>
<td>1</td>
<td>238</td>
<td>163</td>
<td>58</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>126</td>
<td>76</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>69</td>
<td>39</td>
<td>27</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>27</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>38</td>
<td>20</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>24</td>
<td>10</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>13</td>
<td>7</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>777</td>
<td>475</td>
<td>265</td>
<td>37</td>
</tr>
</tbody>
</table>

Post retraction citations and reads

One of the most disturbing issues caused by retracted articles is citations after the article has been retracted (post retraction citations). Bar-Ilan and Halevi (2017) clearly demonstrated for a set of ten articles, that most post retraction citations ignore that the publication has been retracted, and cite the publication as legitimate while supporting the findings of the citing publication.

Here we provide evidence that citations and reads are continued to be accumulated. The number of citations at the date of data collection (October 2014) serves as baseline, as all the articles in the dataset were retracted by then. Citation and reads were collected two more times, in December 2015 and in April 2017. Citation counts were collected from Scopus and reader counts from Mendeley. Table 2 displays the number of citations and reads at each of these data collection points and the increase in percentage between the different data collection points.

Table 2: Growth in the number of citations and reads

<table>
<thead>
<tr>
<th></th>
<th>Citations</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of citations</td>
<td>13,973</td>
<td>15,956</td>
<td>17,385</td>
<td>9,577</td>
<td>13,032</td>
</tr>
<tr>
<td>% growth between</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>consecutive data</td>
<td>14.19%</td>
<td>8.96%</td>
<td>36.08%</td>
<td>53.57%</td>
<td></td>
</tr>
<tr>
<td>points</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% growth between</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>first and last data</td>
<td>24.42%</td>
<td></td>
<td></td>
<td></td>
<td>108.97%</td>
</tr>
<tr>
<td>points</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Within a period of two and a half years, there was an increase of nearly 25% in the number of citations, even though some of the articles have been retracted for many years at the time of
data collection (3 years on average), thus the last data collection point is on average about 5 years after retraction. The number of reads on Mendeley more than doubled within the same time period. It is difficult to imagine that this enormous increase in Mendeley readership is only due to readers being curious as to why the article was retracted. Next, we look at citations and reads per categories (Tables 3 and 4).

Table 3: Post retraction citation growth per category

<table>
<thead>
<tr>
<th>Categories</th>
<th># articles</th>
<th>Sum of cits Oct14</th>
<th>Sum of cits Jan16</th>
<th>Sum of cits Apr17</th>
<th>Ave. cits Oct14</th>
<th>Ave. cits Jan16</th>
<th>Ave. cits Apr17</th>
<th>Growth Oct14 - Apr17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admin. Error</td>
<td>51</td>
<td>166</td>
<td>196</td>
<td>209</td>
<td>3.25</td>
<td>3.84</td>
<td>4.10</td>
<td>0.84</td>
</tr>
<tr>
<td>Ethical misconduct</td>
<td>642</td>
<td>6508</td>
<td>7652</td>
<td>8526</td>
<td>10.14</td>
<td>11.92</td>
<td>13.28</td>
<td>3.14</td>
</tr>
<tr>
<td>Scientific distortion</td>
<td>302</td>
<td>7299</td>
<td>8108</td>
<td>8650</td>
<td>24.17</td>
<td>26.85</td>
<td>28.64</td>
<td>4.47</td>
</tr>
<tr>
<td>Grand Total</td>
<td>995</td>
<td>13973</td>
<td>15956</td>
<td>17385</td>
<td>14.04</td>
<td>16.04</td>
<td>17.47</td>
<td>3.43</td>
</tr>
</tbody>
</table>

Note that even though there are more than twice as many articles in the ethical misconduct category compared to scientific distortion, the articles in the scientific distortion category received more citations than articles in the ethical misconduct category. This is even more emphasized when looking at the average number of citations per paper in each category, and in the average number of new citations received during the two and the half year period. The same is true when considering reads (Table 4). This result is extremely disturbing, because as we noted before, scientific distortion is a hurdle in the advancement of science, and as previous studies showed most of the post retraction citations ignore the fact that these articles were retracted.

Table 4: Post retraction readership growth per category

<table>
<thead>
<tr>
<th>Categories</th>
<th># articles</th>
<th>Sum of reads Oct14</th>
<th>Sum of reads Jan16</th>
<th>Sum of reads Apr17</th>
<th>Ave. reads Oct14</th>
<th>Ave. reads Jan16</th>
<th>Ave. reads Apr17</th>
<th>Growth Oct14 - Apr17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admin. Error</td>
<td>51</td>
<td>201</td>
<td>293</td>
<td>447</td>
<td>5.15</td>
<td>8.88</td>
<td>11.46</td>
<td>6.31</td>
</tr>
<tr>
<td>Ethical misconduct</td>
<td>642</td>
<td>4251</td>
<td>6136</td>
<td>10642</td>
<td>8.05</td>
<td>13.17</td>
<td>18.44</td>
<td>10.39</td>
</tr>
<tr>
<td>Scientific distortion</td>
<td>302</td>
<td>5125</td>
<td>6603</td>
<td>8924</td>
<td>20.02</td>
<td>26.41</td>
<td>32.33</td>
<td>12.31</td>
</tr>
<tr>
<td>Grand Total</td>
<td>995</td>
<td>9577</td>
<td>13032</td>
<td>20013</td>
<td>11.64</td>
<td>17.40</td>
<td>22.44</td>
<td>10.80</td>
</tr>
</tbody>
</table>

Conclusion and recommendations

The aim of this study was to show that retracted articles, and especially those that are retracted because of scientific distortion continue to be cited. The next step will be verifying the assumption that in the majority of the cases, papers are cited positively and their status as retracted is ignored.

One of the possible reasons for the continued citations is that Elsevier’s policy is to have free access to the retracted articles, even though the COPE guidelines require only the retraction notice to be freely available. The cope recommendation is not to delete the article altogether, but to clearly designate the article as retracted with a link or explanation of the reason for retraction. This free access vs. the other articles in the same journal that are behind paywall
may be an incentive for citing. We recommend that Elsevier will change its policy regarding free access to non-open access retracted articles, especially since it seems after conducting a small random check, that SpringerNature and Wiley charge for non-open access retracted articles. In addition, when searching Scopus, often the term “RETRACTED ARTICLE” does not appear in the title of the publication and only when retrieving the full text from the publisher, the searcher becomes aware of the retraction. Of course, there can be other versions of the retracted article freely available on the Web and not designated as retracted, but this is unavoidable.

We are not completely against citing retracted articles, if there is a specific reason for this, and it is clear to the readers that a retracted article is referenced. Publishers routinely run plagiarism checks. It is much easier to discover citations of retracted articles, since probably there are only a few thousand such articles.

Acknowledgments
The first author was supported by EU COST Actions PEERE and KnowEscape.

References
Measuring Model of Collaboration Ability: the Collaborative Rate, Collaborative Breadth and Collaborative Depth

Rongying Zhao¹ Xuqiu Wei² Yang Zhang³

¹ zhaorongying@126.com
² weixuqiu@163.com
³ yangz10@whu.edu.cn
Research Center for Chinese Science Evaluation of Wuhan University, Wuhan (China)
School of information management, Wuhan University, Wuhan (China)

Abstract
It is necessary to measure collaboration ability and identify the most appropriate potential partners in the current research environment. This study proposes a measuring model of collaboration ability, including the collaborative rate, the collaborative breadth and the collaborative depth, to measure author’s collaboration ability, then to identify the most appropriate potential partners. We use the articles published during the period 1978-2016 in Scientometrics for empirical research. Our research shows that collaboration is an important way in the scientific research activities in Scientometrics. Unfortunately, most author’s collaborative breadth and the collaborative depth are lower than mean. Therefore, the author’s scope and stability of collaboration is further strengthened in future. Authors can identify the most appropriate potential partners according to author’s research purpose and the region of the collaborative breadth - the collaborative depth. Therefore, the model can be used to evaluate author’s collaboration ability and identify the most appropriate potential partners.

Keywords measuring mode · collaboration ability · collaborative rate · collaborative breadth · collaborative depth · Scientometrics

Conference Topic
Indicators; Methods and techniques

Introduction
Collaboration is an essential part of academic life. Collaboration has the potential to solve complex scientific problems and promote various political, economic and social agendas (Sonnenwald, 2007). Nowadays, more and more scholars not only pay attention to scientific collaboration, but also carry out a series of studies, such as concept of collaboration (Katz and Martin, 1997), collaboration networks (Zhai, et al., 2014), collaboration pattern (Nankani, et al., 2008), potential of collaboration (Zinsstag, et al., 2005), etc.
How to measure collaboration ability and identify the most appropriate potential partners is the key to carry out collaborative research. In order to further enrich the evaluation system of collaboration, this paper proposes a measuring model of collaboration ability, which contains

Corresponding authors:
Xuqiu Wei(weixuqiu@163.com); Rongying Zhao(zhaorongying@126.com).
the collaborative rate, the collaborative breadth and the collaborative depth, and takes the articles published during the period 1978-2016 in *Scientometrics* as an empirical object to verify the feasibility and operability of the model.

2 Background

With the deepening of scientific research collaboration, it gradually forms a series of evaluation indicators. According to the difference of indicators, the studies are divided into two types of research:

*Single-indicator study:* Single-indicator study refers to the statistical study of single indicator data. For example, Newman (2001) studied a variety of statistical properties of our networks, including numbers of papers written by authors, numbers of authors per paper, numbers of collaborators that scientists have, existence and size of a giant component of connected scientists, and degree of clustering in the networks. Zhai, et al. (2014) studied the evolution and trend of collaboration networks in the field of information systems, and found that average collaboration degree and co-authorship ratio of articles over time were on the rise overall.

*Multi-indicator study:* Multi-indicator study is a comprehensive study of multiple indicator data. Schubert (2012) proposed partnership ability index. It combined the number of co-authors and the times each of them acted as co-authors with a given author exactly the same way as Hirsch’s h-index combines the number of publications and their citation rate. Wang, et al. (2014) introduced Schubert’s partnership ability index into China and analyzed its own unique theoretical and practical application value. And Sabaghinejad, et al. (2016) estimated the partnership ability of *Scientometrics* journal authors based on Web of Science from 2001 to 2013 via this index.

Single-indicator study seems relatively simple. Although multi-indicator study (partnership ability index) combined with a variety of indicator data seems relatively complex, it ignores the maximum scope of collaboration and the maximum collaboration intensity. The maximum scope of collaboration and the maximum collaboration intensity are important manifestation of authors’ collaboration ability. Therefore, this study combines the advantages of single-indicator study and multi-indicator study, proposes a measuring model of collaboration ability, which contains the collaborative rate, the collaborative breadth and the collaborative depth at author’s level, and uses the collaborative breadth and the collaborative depth to identify the most appropriate potential partners.

3 Measuring model / indicators/ data collection

*Measuring model*

The measuring model of collaboration ability is proposed in this section. The data processing of measuring model (Shown in Figure 1) is as followed: firstly, author fields in articles are extracted by using Bibexcel software to obtain OUT1 file and CIT1 file (OUT1 file is the distribution of author field per article, while CIT1 file is the cumulative frequency of author field). And author fields in collaborative articles are extracted to obtain OUT2 file and CIT2 file. Secondly, from CIT1 file, it will get the number of articles per authors; from CIT2 file, it will get the number of collaborative articles of this author, then it will calculate the collaborative rate (author level). And finally, the processed of OUT2 file through VBA and Excel obtains the collaborative breadth and the collaborative depth.
In this section, the collaborative rate, the collaborative breadth and the collaborative depth will be defined:

**Collaborative rate:** In 2012, Wei and Ji (2012) proposed the collaborative rate at country’s or institute’s level. It was determined by the percentage of collaborative publications relative to total publications for each country or institute. According this, the author’s collaborative rate is determined in this paper. It is that the percentage of collaborative literatures relative to total literatures for each author. It is used to reflect the author’s collaborative willingness in the discipline or research field. The larger value it is, the higher collaborative willingness the author has. In other word, it is easy for one author to build partnerships with the author who has high collaborative rate.

**Collaborative breadth:** It is used to reflect the author’s collaboration and communication ability, the scope of collaboration, namely, the total number of authors is that the author can be established collaboration relationship in a discipline or research field. For example, author a1, author a2 and author a3 published paper p1 together, author a1 and author a2 published paper p2 together, and author a1 and author a4 published paper p3, then the collaborative breadth of author a1 is 3 (author a2, author a3 and author a4), the collaborative breadth of author a2 is 2 (author a1 and author a3), the collaborative breadth of author a3 is 2 (author a1 and author a2), while the collaborative breadth of author a4 is 1 (author a1). Therefore, the collaboration and communication ability of author a1 is stronger than that of author a2, author a3 and author a4.

**Collaborative depth:** It is used to reflect the depth of author’s collaboration and communication, the stability degree of collaboration. It can use the maximum collaborative number between the author and other authors to represent. In the above example, author a1 and author a2 published 2 papers, author a1 and author a3 published 1 paper, and author a1 and author a4 co-published 1 paper, then the collaborative depth of both author a1 and author a2 is 2, the collaborative depth of author a3 and author a4 is 1. That is to say, one author can collaborate with author a1 or author a2 two times, however, he/she can only collaborate with author a3 or author a4 one time.

**Collaborative breadth and Collaborative depth:** In order to comprehensively reflect authors’ collaboration ability, then to identify the most appropriate potential partners, in this study, X-axis presents the collaborative breadth and Y-axis presents the collaborative depth to draw the
collaborative breadth - the collaborative depth scatter plot (Shown in Figure 2). To conveniently analyze the collaborative breadth and the collaborative depth of author’s, it should standard the above two indexes, namely, it scales the data, so that it will fall within a particular range, such as [0,1]:

\[
\text{Collaborative breadth (relative value)} = \frac{\text{Collaborative breadth (absolute value)}}{\text{Max collaborative breadth (absolute value)}} \tag{1}
\]

\[
\text{Collaborative depth (relative value)} = \frac{\text{Collaborative depth (absolute value)}}{\text{Max collaborative depth (absolute value)}} \tag{2}
\]

![Collaborative breadth - the collaborative depth scatter plot (Simple)](image)

**Figure 2**  The collaborative breadth - the collaborative depth scatter plot (Simple)

To visually show the author’s position in the scatter plot and achieve the purpose of identifying potential partners, two reference lines (the mean of the collaborative breadth and the mean of collaborative depth) are delineated in Figure 2. In Figure 2, the assumed value of the collaborative breadth or the collaborative depth is 1. Hence, the scatter plot is divided into four regions, named region I, region II, region III and region IV.

**Region I**: Both the collaborative breadth and the collaborative depth are higher. This indicates the author’s scope of collaboration is wider, and the stability degree of collaboration is higher. Therefore, the authors in this region are the best potential partners. Namely, the authors have more opportunity to establish collaboration relationship with other large authors, besides the stability degree of collaboration is also high.

**Region II**: The collaborative breadth is lower, while the collaborative depth is higher. It shows that the author’s scope of collaboration is limited, but the stability degree of collaboration is also relatively higher. Therefore, the authors in this region focus on the stability degree of collaboration.

**Region III**: Both the collaborative breadth and the collaborative depth are lower. This describes that the author’s scope of collaboration is limited and the stability degree of collaboration is also relatively low.

**Region IV**: The collaborative breadth is higher, while the collaborative depth is lower. It says that the author’s scope of collaboration is higher, but the stability degree of collaboration is also relatively low. Therefore, the authors in this region focus on the collaboration scope.

From author’s regional location shown in Figure 2, it can be learned that: the authors who have wider collaboration scope or have higher stability degree of collaboration provides a certain reference for identifying potential partners.
Data collection

Scientometrics is concerned with the quantitative features and characteristics of science and scientific research to provide guidance for the effective conduct of scientific research activities. During the developmental period of Scientometrics, the foundation of the journal *Scientometrics* (in September, 1978) was a landmark event. Following the works of previous researchers (Chen, et al., 2013; Hou, et al., 2008; Zhao, et al., 2015), this study used the journal as a representative model of Scientometrics research. Data was acquired from the Web of Science in Feb 2017. All 4,242 articles-covering articles and proceedings papers-published in the journal of *Scientometrics* in 1978–2016 (publication year) were downloaded. Among the above 4,242 articles, 2,890 articles were published by 2 authors or more authors.

4 Results and discussions

The importance of collaboration at article level

Table 1 lists the number of articles, number of collaborative articles and collaborative rate at articles level per year.

In 1978, there are 5 articles published in Scientometrics. However, in 2016, there are 341 articles published in Scientometrics. Compared the number of articles in 1978, the number of articles has increased in 2016. To further illustrate this growth trend, the trend line is developed using publication year and number of articles as variables. The trend line of publication year and number of articles is that \( y=17.565e^{0.0738x} \), \( R^2=0.8742 \). The increase in the number of articles suggests the importance of Scientometrics in scientific research management and policy making.

Table 1  Distribution of the number of articles, number of collaborative articles and collaborative rate at articles level per year

<table>
<thead>
<tr>
<th>Year</th>
<th>NA</th>
<th>NCA</th>
<th>CR</th>
<th>Year</th>
<th>NA</th>
<th>NCA</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1978</td>
<td>5</td>
<td>3</td>
<td>60.00%</td>
<td>1998</td>
<td>82</td>
<td>43</td>
<td>52.44%</td>
</tr>
<tr>
<td>1979</td>
<td>18</td>
<td>14</td>
<td>77.78%</td>
<td>1999</td>
<td>126</td>
<td>58</td>
<td>46.03%</td>
</tr>
<tr>
<td>1980</td>
<td>32</td>
<td>8</td>
<td>25.00%</td>
<td>2000</td>
<td>82</td>
<td>47</td>
<td>57.32%</td>
</tr>
<tr>
<td>1981</td>
<td>29</td>
<td>11</td>
<td>37.93%</td>
<td>2001</td>
<td>89</td>
<td>50</td>
<td>56.18%</td>
</tr>
<tr>
<td>1982</td>
<td>24</td>
<td>13</td>
<td>54.17%</td>
<td>2002</td>
<td>84</td>
<td>51</td>
<td>60.71%</td>
</tr>
<tr>
<td>1983</td>
<td>22</td>
<td>10</td>
<td>45.45%</td>
<td>2003</td>
<td>81</td>
<td>61</td>
<td>75.31%</td>
</tr>
<tr>
<td>1984</td>
<td>24</td>
<td>7</td>
<td>29.17%</td>
<td>2004</td>
<td>86</td>
<td>59</td>
<td>68.60%</td>
</tr>
<tr>
<td>1985</td>
<td>48</td>
<td>22</td>
<td>45.83%</td>
<td>2005</td>
<td>99</td>
<td>64</td>
<td>64.65%</td>
</tr>
<tr>
<td>1986</td>
<td>39</td>
<td>19</td>
<td>48.72%</td>
<td>2006</td>
<td>149</td>
<td>102</td>
<td>68.46%</td>
</tr>
<tr>
<td>1987</td>
<td>45</td>
<td>21</td>
<td>46.67%</td>
<td>2007</td>
<td>129</td>
<td>98</td>
<td>75.97%</td>
</tr>
<tr>
<td>1988</td>
<td>48</td>
<td>23</td>
<td>47.92%</td>
<td>2008</td>
<td>128</td>
<td>90</td>
<td>70.31%</td>
</tr>
<tr>
<td>1989</td>
<td>64</td>
<td>31</td>
<td>48.44%</td>
<td>2009</td>
<td>187</td>
<td>135</td>
<td>72.19%</td>
</tr>
<tr>
<td>1990</td>
<td>61</td>
<td>26</td>
<td>42.62%</td>
<td>2010</td>
<td>223</td>
<td>163</td>
<td>73.09%</td>
</tr>
<tr>
<td>1991</td>
<td>78</td>
<td>38</td>
<td>48.72%</td>
<td>2011</td>
<td>215</td>
<td>162</td>
<td>75.35%</td>
</tr>
<tr>
<td>1992</td>
<td>75</td>
<td>37</td>
<td>49.33%</td>
<td>2012</td>
<td>252</td>
<td>188</td>
<td>74.00%</td>
</tr>
<tr>
<td>1993</td>
<td>57</td>
<td>34</td>
<td>59.65%</td>
<td>2013</td>
<td>249</td>
<td>194</td>
<td>77.91%</td>
</tr>
<tr>
<td>1994</td>
<td>58</td>
<td>28</td>
<td>48.28%</td>
<td>2014</td>
<td>338</td>
<td>283</td>
<td>83.73%</td>
</tr>
<tr>
<td>1995</td>
<td>70</td>
<td>36</td>
<td>51.43%</td>
<td>2015</td>
<td>343</td>
<td>285</td>
<td>83.09%</td>
</tr>
<tr>
<td>1996</td>
<td>87</td>
<td>45</td>
<td>51.72%</td>
<td>2016</td>
<td>341</td>
<td>291</td>
<td>85.34%</td>
</tr>
<tr>
<td>1997</td>
<td>75</td>
<td>40</td>
<td>53.33%</td>
<td>Total</td>
<td>4,242</td>
<td>2,890</td>
<td>68.13%</td>
</tr>
</tbody>
</table>

Description: NA presents the number of articles, NCA presents the number of collaborative articles and CR presents the collaborative rate at article level.

Although the number of collaborative articles also shows growth trend (Trend line: \( y=6.9051e^{0.0922x} \), \( R^2=0.9296 \)), the importance of collaboration in Scientometrics at article level cannot be judged. Therefore, the collaborative rate at articles level is calculated to illustrate this point. In 1987, it is 60.00%, less than than mean (68.13%); while in 2016, it ups to 85.34%
more than the mean (68.13%). Generally speaking, the collaborative rate at articles level also has increased. To some extent, it says collaboration is important in the research of Scientometrics. To better illustrating the importance of collaboration in Scientometrics at article level, the curve estimation was developed using publication year and collaborative rate at article level as variables. The result (shown in Table 2) indicates that it relatively accords with the linear model. It can be seen that with the development of science and technology, scientific research cooperation becomes more and more important.

This phenomenon occurs not only in Scientometrics, but also occurs in other research areas. For example, Gazni, et al. (2012) found that Collaborative rates continued to grow across disciplines and countries. Gazni, et al., (2012) found that the collaborative rate of the world had risen from 15.02% in 2000 to 16.85% in 2005 to 20.09% in 2010; while American collaborative rate had risen from 20.57% in 2000 to 23.63% in 2005 to 30.28% in 2010. Wei and Li (2014) found that the collaborative rate (article level) had been increased from 34.81% in 2004 to 54.36% in 2013 in library and information science. Therefore, all above indicate the importance of scientific research collaboration.

**Table 2  Model Summary and Parameter Estimates**

<table>
<thead>
<tr>
<th>Equation</th>
<th>Model Summary</th>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>R Square</td>
<td>F</td>
<td>df1</td>
</tr>
<tr>
<td>Linear</td>
<td>.600</td>
<td>55.497</td>
</tr>
</tbody>
</table>

The independent variable is Year.

**Author’s collaborative willingness: collaborative rate**

The above section reflects the importance of collaboration at articles level. Author’s collaborative willingness (collaborative rate) determines the author's attitude towards scientific research collaboration. By the statistic, 5,296 authors are discovered from 4,242 articles in 1978-2016 in Scientometrics. The number of articles, number of collaborative articles and collaborative rate at author level are shown in Table 3.

**Table 3  The number of articles, number of collaborative articles and collaborative rate at author level**

<table>
<thead>
<tr>
<th>Author</th>
<th>NA</th>
<th>NCA</th>
<th>CR</th>
<th>Author</th>
<th>NA</th>
<th>NCA</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gnanzel, W</td>
<td>113</td>
<td>99</td>
<td>0.88</td>
<td>Lewison, G</td>
<td>29</td>
<td>18</td>
<td>0.62</td>
</tr>
<tr>
<td>Schubert, A</td>
<td>76</td>
<td>67</td>
<td>0.88</td>
<td>Ho, YS</td>
<td>28</td>
<td>26</td>
<td>0.93</td>
</tr>
<tr>
<td>Leydesdorff, L</td>
<td>63</td>
<td>39</td>
<td>0.62</td>
<td>Kretschmer, H</td>
<td>25</td>
<td>16</td>
<td>0.64</td>
</tr>
<tr>
<td>Rousseau, R</td>
<td>60</td>
<td>51</td>
<td>0.85</td>
<td>Chen, DZ</td>
<td>24</td>
<td>24</td>
<td>1.00</td>
</tr>
<tr>
<td>Braun, T</td>
<td>55</td>
<td>53</td>
<td>0.96</td>
<td>Vinkler, P</td>
<td>24</td>
<td>20</td>
<td>0.00</td>
</tr>
<tr>
<td>Egghe, L</td>
<td>51</td>
<td>14</td>
<td>0.27</td>
<td>Meyer, M</td>
<td>24</td>
<td>18</td>
<td>0.75</td>
</tr>
<tr>
<td>Bornmann, L</td>
<td>38</td>
<td>35</td>
<td>0.92</td>
<td>Van Raan, AFJ</td>
<td>23</td>
<td>13</td>
<td>0.57</td>
</tr>
<tr>
<td>Thedwall, M</td>
<td>37</td>
<td>35</td>
<td>0.95</td>
<td>Courial, JP</td>
<td>22</td>
<td>19</td>
<td>0.86</td>
</tr>
<tr>
<td>Moed, HF</td>
<td>37</td>
<td>30</td>
<td>0.81</td>
<td>Park, HW</td>
<td>23</td>
<td>22</td>
<td>0.96</td>
</tr>
<tr>
<td>Gupta, BM</td>
<td>33</td>
<td>31</td>
<td>0.94</td>
<td>Garg, KC</td>
<td>22</td>
<td>21</td>
<td>0.95</td>
</tr>
<tr>
<td>Huang, MH</td>
<td>32</td>
<td>32</td>
<td>1.00</td>
<td>Van Leeuwen, TN</td>
<td>21</td>
<td>20</td>
<td>0.95</td>
</tr>
<tr>
<td>Abrome, G</td>
<td>30</td>
<td>30</td>
<td>1.00</td>
<td>Debackere, K</td>
<td>21</td>
<td>21</td>
<td>1.00</td>
</tr>
<tr>
<td>D'Angelo, CA</td>
<td>30</td>
<td>30</td>
<td>1.00</td>
<td>Zitt, M</td>
<td>21</td>
<td>17</td>
<td>0.81</td>
</tr>
<tr>
<td>Thijs, B</td>
<td>29</td>
<td>29</td>
<td>1.00</td>
<td>Persson, O</td>
<td>20</td>
<td>16</td>
<td>0.80</td>
</tr>
</tbody>
</table>
| Guan, JC   | 29 | 29  | 1.00| ...        | ...| ... | ...

Description: NA presents the number of articles, NCA presents the number of collaborative articles and CR presents the collaborative rate at author level.
As shown in Table 3, among those 5,296 authors, Glanzel, Schubert, Leydesdorff, Debackere, Zitt, Persson, Bordons and other 21 authors published more than 20 articles in 2006-2016. Among the 28 authors, except for author Eghe, 27 authors have high collaborative rate. Here, it will have a question that if author whose number of collaborative articles is high, his/her number of articles is high. Based on this problems, the correlation between number of collaborative articles and number of articles is analyzed by SPSS20, shown in Table 4.

### Table 4  Correlations between number of collaborative articles and number of articles at author level

<table>
<thead>
<tr>
<th>Number of articles</th>
<th>Pearson Correlation</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.954*</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>5296</td>
<td>5296</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).

From Table 4, the Pearson Correlation of between number of collaborative articles and number of articles is 0.954. This illustrates that there is a strong positive correlation between the two, and the probability P-value of the correlation coefficient test is approximately 0. Therefore, when the Correlation is significant at the 0.01 level, it should reject the original hypothesis of correlation coefficient test, that two of the total is not zero correlation. This further illustrates the importance of collaboration in Scientometrics.

In order to illustrate author’s collaborative willingness, the collaboration rate of 5,296 authors is measured. The Statistics of author’s collaborative rate is shown in Table 5.

### Table 5  Statistics of author’s collaborative rate

<table>
<thead>
<tr>
<th>N</th>
<th>Valid</th>
<th>5296</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>.8887</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>Mode</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>0.29454</td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Percentiles</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.5000</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1.0000</td>
<td></td>
</tr>
</tbody>
</table>

From Table 5, the Mean of 5,296 authors’ collaborative rate is 0.8887, the Median is 1, the Mode is 1 and the Std. Deviation is 0.29454. To some extent, those indicators show that authors in Scientometrics have high collaborative willingness. At the same time, Table 5 shows the Minimum is 0, i.e. some literatures were published only by one author. However, from the Percentiles in Table 5, only 15% authors, whose collaborative rate are less than 1, 85% authors, whose collaborative willingness are 1. It illustrates that authors in Scientometrics are willing to collaborate with other authors. This is to lay the foundation for potential collaboration in Scientometrics. At the same time, this also further suggests the importance of collaboration in research activities.

**Author’s collaboration and communication ability: collaborative breadth**

To reflect the author’s collaboration and communication ability, the scopes of collaboration, in Scientometrics, the collaborative breadth is introduced. By the statistic, 4,830 authors, who publish articles with other author(s), are discovered from 2,890 articles during the period 1978-2016 in Scientometrics. The collaborative breadth and its relative value of 4,830 authors are calculated, shown in Table 6. From Table 6, we can learn that the highest collaborative breadth is Rousseau, whose value is 60, i.e. author Rousseau has 60 authors to establish collaboration relationship during the period 1978-2016 in Scientometrics. The frequency distribution of
collaborative breadth is illustrated in Figure 3.

Figure 3 The frequency distribution of collaborative breadth

From Figure 3, the Mean of 4,830 authors’ collaborative breadth is 3.6609, the Median is 3, the Mode is 2, the Std. Deviation is 3.65641, the Minimum is 1, and the Maximum is 60. Besides, the collaborative breadth of 4,830 authors are mainly in the interval [1, 5]. This illustrates that authors in Scientometrics have 1-5 collaborators.

The main factors affecting the collaborative breadth are as followed:

(1) Researchers’ qualification. At universities or research institutions, the researchers (such as doctoral supervisor, master supervisor) always have higher qualifications, and they also have wider collaboration scope. 

① Researchers as doctoral supervisors or master supervisors are responsible for guiding students to familiar with the theoretical knowledge and cultivate the ability of students to identify problems, analyze problems and solve problems. During this process, the researchers have potential partners.

② Since those researchers with higher qualification, some researchers whose qualification is relatively low are willing to carry out collaborative research with them, then to enhance their own influence and research level.

(2) Researchers’ ability. To some extent, the collaborative breadth is determined by researchers’
ability. In process of research activities, researchers are willing to cooperate with the researchers who have higher capacity. Through the collaborative research, researchers who have lower ability can keep abreast of their lack in scientific research, and get some lessons and skills to quickly enhance their research abilities.

(3) Distance between researchers. The collaborative breadth is also affected by the distance between researchers. There are two kinds of distance: ① Spatial distance. In general, it is easy to see the collaboration between the same institutions. With the development of technology and internet, this phenomenon is eased. It is common to see multi-agency collaboration, inter-regional collaboration and global collaboration nowadays. And the multi-agency collaboration, inter-regional collaboration and global collaboration are gradually becoming new researches. ② Discipline difference. To a certain extent, the difference of discipline knowledge also limits the collaboration between authors: communication and collaboration is very common in the same discipline, the interdisciplinary communication and collaboration between the authors is less. However, knowledge of multiple disciplines is needed to solve many practical problems (Xu, et al., 2016). This promotes the cross combination of theories, methods, and technologies among different disciplines.

Author’s the depth of collaboration and communication: collaborative depth
To reflect the author’s depth of collaboration and communication, the stability degree of collaboration, in Scientometrics, the collaborative breadth is introduced. The collaborative depth and its relative value of 4,830 authors are calculated, shown in Table 7. From Table 7, the highest collaborative depth is author Braun and Schubert, whose value is 40. In 1981, author Braun and Schubert firstly published Some Scientometric Measures of Publishing Performance for 85 Hungarian Research Institutes together in Scientometrics. Over the next 22 years, they published 39 articles together. Obviously, their depth of collaboration and communication is deepest, and their relationship is most closely related. The frequency distribution of collaborative depth is illustrated in Figure 4.

![Collaborative depth](image1)
![Collaborative breadth](image2)

**Figure 4** The frequency distribution of collaborative depth

**Figure 5** The collaborative breadth - the collaborative depth scatter plot (982 authors)

From Figure 4, the Mean of 4,830 authors’ collaborative depth is 1.4290, the Median is 1, the Mode is 1, the Std. Deviation is 1.64461, the Minimum is 1, and the Maximum is 40. Besides, the collaborative depth of 4,830 authors are mainly 1. This illustrates that the stability degree...
of collaboration of authors in Scientometrics are 1. The main factors affecting the collaborative depth are as followed:

(1) **Collaborative time.** If two researchers have longer collaborative time, they will have more opportunity to carry out scientific research activities. Thus, the collaborative depth of them is higher. We believe that there are two time factors impact author’s collaborative depth. One is educational system, especially for supervisor or students. In China, the educational system of academic master’s education in library science is 2~3 years (Jiang and Yang, 2016); in the USA, it is 3~4 years; in England, it is 1~2 or 1~3 years (Wang, et al., 2009). This greatly limits teacher or students carry out researches together, then limits teachers’ or student’s collaborative depth. The other is time lag of paper publishing (Du, 2004). It takes a certain of amount of time for academic from received to be published. Unfortunately, in this study, the unpublished papers are not included in the measurement of collaborative depth. To a certain extent, time lag of paper publishing also affects author’s collaborative depth. However, this factor is difficult to avoid in the actual calculation process.

(2) **Research capacity level of bilateral collaboration.** From the definition of collaborative depth, the value of the collaborative depth is depended on both researchers. When both researchers have a strong research capacity, the collaborative depth is relatively high; when one side’s is high and the other’s is relatively weak, the collaborative depth is determined by the side whose research capacity is weak; and when both two researchers’ research capabilities are relatively weak, the collaborative depth is relatively weak. Therefore, we can say that the collaborative depth is determined by the research whose research capacity is relatively weak.

**Collaborative breadth & collaborative depth**

Either the collaborative breadth or the collaborative depth only reflects collaboration ability from a single aspect. To comprehensively reflect author’s collaboration ability in Scientometrics during the period 1978-2016, the collaborative breadth (relative) is calculated according to the formula (1), while the collaborative depth (relative) is calculated according to the formula (2). And the collaborative breadth - the collaborative depth scatter plot is drawn by SPSS20. Since labels exceed 1,000 limit in SPSS20, random sample (approximately 20% of all cases) is selected via SPSS20. In the end, we only show the collaborative breadth and collaborative depth of 982 authors in the scatter plot, shown in Figure 5.

Among 4,830 authors, in region Ⅰ, there are 646 authors (accounting 13.37%); in region Ⅱ, there are 370 authors (accounting 7.66%), in region Ⅲ, there are 2,756 authors (accounting 57.06%); and in region Ⅳ, it has 1,058 authors (accounting 21.90%). From the above result, it illustrates that the most of researchers (accounting 35.28%) during the period 1978-2016 in Scientometrics have higher collaborative breadth. And 21.03% of authors have higher collaborative depth. Therefore, the author’s scope and stability degree of collaboration is further strengthened.

The author’s position in Figure 5 shows the collaboration ability during the period 1978-2016 in Scientometrics. From the characteristics of author’s distribution in scatter plot, it is helpful for researchers, especially new researchers, to identify the most appropriate potential partners: (1) “Best” regional collaboration. In this region (region Ⅰ), both the collaborative breadth and the collaborative depth are higher, i.e. the author’s scope of collaboration is wider and the stability degree of collaboration is higher. The collaboration with authors in region Ⅰ is not only beneficial to expanding the scope of collaboration, but also help to carry out in-depth collaboration to improve their academic influence together in Scientometrics.
(2) Deepen stability regional collaboration. In this region (region Ⅱ), the collaborative breadth is lower than the mean, while the collaborative depth is higher than the mean. Compared with region Ⅰ, the scope of collaboration is limited, but the stability degree of collaboration is higher. Therefore, the collaboration with authors in region Ⅱ is help to carry out depth collaboration to mine some research contents or subjects deeply.

(3) General regional collaboration. In this region (region Ⅲ), both the collaborative breadth and the collaborative depth are lower. Researchers in this region should increase their scope of collaboration and stability degree of collaboration. So, they can collaborate with authors in region Ⅰ, region Ⅱ and region Ⅳ.

(4) Expand scope regional collaboration. In this region (region Ⅳ), the collaborative breadth is higher than the mean, while the collaborative depth is lower than the mean. Compared with region Ⅰ, the scope of collaboration is higher, while the stability degree of collaborative is limited. Therefore, the collaboration with authors in region Ⅳ is help to increase the scope of collaboration.

In short, according to author’s research purpose and the region of the collaborative breadth - the collaborative depth, he/she can identify the most appropriate potential partners to carry out collaborative research, especially for the authors who are new researchers in this field or whose collaborative breadth or collaborative depth should be further strengthened.

5 Conclusions

In scientific research, collaboration is becoming more and more important. It is important to measure author’s collaboration ability and identify the most appropriate potential partners. After research, we find that: (1) with the development of Scientometrics, both the number of articles and the number of collaborative articles have increased. What’s more, the collaborative rate at article level has increased. (2) The analysis of author’s collaborative rate, it find that author’s collaborative willingness are higher in Scientometrics. This not only further shows the importance of collaboration, but also reflects the author’s views on collaborative behavior. (3) The collaborative breadth of 4,830 authors are mainly in the interval [1,5], while the maximum is 60; the stability degree of collaboration of authors in Scientometrics are 1, while the maximum is 40. The author’s scope and stability degree of collaboration is further strengthened. And (4) the scatter plot is divided into 4 regions. According to the actual situation of the author, he/she can identify the most appropriate potential partners.

This research contributes several perspectives. First, this study proposes measuring model of collaboration ability. The collaborative rate reflects author’s collaborative willingness, however, the collaborative breadth and collaborative depth reflects the degree of collaboration and communication. Second, to reflect authors’ collaboration ability from multi-indictor data, this study comprehensively analyzes author’s collaborative breadth and collaborative depth in a scatter plot. Besides, based on the scatter plot, researcher can identify the most appropriate potential partners for himself/herself. And finally, in this article, we use empirical method to measure author’s collaboration ability and identify the most appropriate potential partners. It will provide an example for other disciplines or research field.

In this study, the present study has a little less, to be further improved: the empirical analysis only takes Scientometrics as a case, interdisciplinary collaboration ability will be discussed in the future. Overall, this study proposes measuring model of collaboration ability to measure author’s collaboration ability and identify the most appropriate potential partners.
Acknowledgments
This work is supported by National Social Science Foundation in China (Grant No.16BTQ055).

References:
Some Reflections to China’s International Collaboration

Yi Han¹, Yanxiao Liu², Lanni Shen³ and Bihui Jin⁴

¹hanyi72@swu.edu.cn, ²1208913462@qq.com, ³1837495883@qq.com

¹²³College of Computer and Information Science, Southwest University(P R China)

⁴jinbh@mail.las.ac.cn

⁴National Science Library, Chinese Academy of Sciences(P R China)

Abstract
This paper is to reflect the China’s international collaboration in science during the period 1995-2015 and to discover the dynamics and the channel of international collaboration between China and the main scientific and technological powers. USA, Japan, Germany, UK, and France are chosen the international target countries to compare, the collaborative data of Big 6 are retrieved in Web of Science, and some descriptive statistical analyses are given. The data show that China’s share of internationally co-authored papers is still considerably lower than that of the other main countries. The numbers of collaborative papers in five nations have grown much faster than the numbers of non-collaborative ones. In China collaborative papers grow only slightly faster than non-collaborative ones, especially over the latest years. China’s international collaboration in science has obvious disciplinary preferences, focusing mainly on traditional fundamental disciplines. Overseas Chinese play an extremely important role in mainland China’s international collaboration, providing an optimal channel to overcome linguistic and cultural barriers.

Keywords
International scientific and technological collaboration; China’s international collaboration; overseas Chinese phenomenon

Conference Topic
Country-level studies

Introduction
By the end of the fifties in the 20th century, Smith (1958) noted a trend towards more and more multiple authorship. This led Price (1963) to predict the demise of the single-author article. Although this prediction has not come true, the number of single-author articles has indeed steadily declined. Wuchty et al. (2007) confirmed the general increase of the influence of teams in science, including that by this phenomenon the process of knowledge creation has fundamentally changed over the years. Scientific collaboration can be described as a process by which scientists with common interests work together to create new knowledge (Katz, 1994, 1997; Coccia & Wang, 2016). It is generally accepted that team work increases quality and productivity and hence the received number of citations, although collaboration does not necessarily lead to success, such as a lot of citations.(Figg et al, 2006; Glänzel, 2008; Elango et al,2015). Generally speaking, nowadays a scientist is not anymore an isolated actor but one member, among many, of an international research team, working together to discover the rules of nature and society (Cronin, 1982; Finardi & Buratti,2016). Due to differences in research accumulation during investigations there
are always leading teams and followers (Zhou & Bormman, 2015; Ma et al., 2016). If researchers collaborate with the leaders, knowledge and technical barriers are reduced and this helps them to reach the research front. Yet, leaders too may benefit from collaborating, as increased manpower can greatly accelerate the development and diffusion of new knowledge, new technologies and new methods.

Developing countries can greatly extend their scientific productivity and strength, in the same time increasing the visibility of their researchers, by collaborating with leading countries (Rigby & Edler, 2005; Jonkers & Tijssen, 2008; Zhou et al., 2009; Ynalvez & Shrum, 2011; Wagner et al., 2015). Collaboration gives, moreover, direct access to complementary knowledge or skills, unique sites and facilities, and shares costs and risks (Kim, 2005; Birnholtz, 2007; Haustein et al., 2011). Starting as a developing (or at least a not-yet-fully-developed) nation China’s science and technology has increased considerably during the latest twenty years. The exponential increase in the absolute and relative number of WoS papers bears testimony of this fact (Jin & Rousseau, 2005). Yet, outstanding scientific and technological discoveries are not distributed evenly over countries, regions or provinces (or states) within a country (Wang et al., 2005; Zhou et al., 2009). By studying international collaboration we intend to obtain a better insight in the position of China with respect to other big countries.

We would like to add one more point, namely that collaborative research may not only lead to publications but also to patents (Chen et al., 2013). This aspect is not the topic of this paper. Hence, the main purpose of this article is to describe China’s position in the international fray, restricted to its output in scientific papers.

Data and Methods

The articles jointly written by several authors reflect scientific and technological collaboration. Consequently international collaboration is reflected by articles jointly published by authors in different countries. Articles co-authored by Chinese scientists in the context of international collaboration are mainly published in international journals. Zheng et al. (2012) mentioned that in domestic journals only 1.5% of the publications are the result of international collaboration. Hence, if one wants to study China’s international collaboration as shown by published articles one must make use of an international database. Consequently we use Thomson Reuter’s Web of Science (WoS). Publications are restricted to the article and review type.

The United States, the UK (taking care of the fact that the UK consists of England, Scotland, Wales and Northern Ireland), Germany, Japan and France are, together with China the countries producing the most articles. They will be referred to as the Big 6. In 2015, the total articles’ number recorded in WoS is 2,632,093, and the number produced by the Big 6 is 1,529,243. Nowadays we focus on these six countries. Taking the names of the chosen countries as retrieval terms, the retrieved data are divided into two groups: international collaborative papers on the one hand and non-international collaboration on the other (authors are all from China, including single author papers).
By choosing the above mentioned countries we expect to grasp the main trends and characteristics of international collaboration with special emphasis on the role played by China. Furthermore, we explore the dynamics of international collaboration and try to find the optimal path for China’s scientists when it comes to participation in international scientific and technological endeavors. Occasionally we will refer to papers written as the result of an international collaboration as “co-papers” (as a short-hand).

Because of this aim, the WoS data set of articles and reviews in 1995-2015 are retrieved. Besides all papers (of article and review type), we also consider separately the collaborative papers, written in collaboration with the main countries with which these countries collaborate the most. Articles by overseas Chinese are especially collected as well as the fields in which collaboration occurs. Data were collected in September 2016.

Results

International collaboration trends of the BIG 6 countries

Taking the name of Big 6 as retrieval terms, the retrieved data are divided into two groups: international collaborative papers, of which the affiliation consists of two or more than two countries, and non-international collaborative one, of which the affiliation only one country. The full counting is recorded every one year since 1995, and the growth rate is calculated respectively. The calculation results are shown in Table 1.

<table>
<thead>
<tr>
<th>Year</th>
<th>Non-collaborative papers</th>
<th>Co-papers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>USA</td>
<td>CHINA</td>
</tr>
<tr>
<td>1997</td>
<td>-3%</td>
<td>26%</td>
</tr>
<tr>
<td>1999</td>
<td>25%</td>
<td>48%</td>
</tr>
<tr>
<td>2001</td>
<td>0%</td>
<td>51%</td>
</tr>
<tr>
<td>2003</td>
<td>26%</td>
<td>59%</td>
</tr>
<tr>
<td>2005</td>
<td>-4%</td>
<td>46%</td>
</tr>
<tr>
<td>2007</td>
<td>3%</td>
<td>28%</td>
</tr>
<tr>
<td>2009</td>
<td>8%</td>
<td>26%</td>
</tr>
<tr>
<td>2011</td>
<td>3%</td>
<td>22%</td>
</tr>
<tr>
<td>2013</td>
<td>4%</td>
<td>32%</td>
</tr>
<tr>
<td>2015</td>
<td>-3%</td>
<td>24%</td>
</tr>
</tbody>
</table>
The data in Table 1 show that, except for China, the growth rate of international collaborative papers is much faster than the growth rate of non-collaborative ones. What could be the reason for this phenomenon?

When a society matures, the division of labor and in particular the numbers of practitioners engaged in scientific and technological research may stay constant, leading to a stabilization of its scientific and technological output. In order to maintain their international position developed countries must adapt their scientific policies and try to increase their attractiveness for foreign scientists, heavyweights as well as average ones. These and other factors must be optimized to maximize scientific and technological output. Supported by a national R&D policies, international collaboration promotes scientific development and progress, leading to prosperity in the collaborating nations. For these reasons developed states have vigorously strengthened international collaboration over the past twenty years. Promoting scientific and technological collaboration has become an important point in the research policies of most countries all over the world.

In contrast, China is not yet a fully developed country, and its society is changing at full speed. On the one hand, the number of practitioners engaged in scientific and technological activities in China continues to increase, and the number of scientific papers maintains a sustained high growth. On the other hand, there is no doubt that it is a good choice to promote scientific and technological development through international collaboration. Therefore, all Chinese papers (internationally collaborated or not) maintained a double-digit growth during the recent 20 years. Furthermore, while China's growth rate of non-collaborative papers tends to decrease after the peak in 2003; its international collaborative papers kept a sustained strong increase, and its growth rate is also higher than that of the other five nations. But until 2015, China’s scientific output(330,927) is still only about half of that of USA(646,528). As a not-yet-fully-developed country, it is of great importance for China to develop a sound scientific policies and to promote scientific research, including international collaboration, to follow or even lead world trends.

The role played by China in international collaboration

In international collaboration, the ranks of authors in the byline of papers usually reflect the contribution of participants, reflecting role differences in collaboration. In particular, the corresponding author is often the person responsible for the research project. Therefore, by analyzing whether an author in collaborative papers is the corresponding author or not, the scientific contribution of authors in such papers can be analyzed. In this section, we investigate if China undertook either a leading role or
a subordinate one in international collaboration by using the available information on corresponding authors. The proportion of corresponding authors, written as CAs, is calculated, and the publishing years are selected randomly.

Table 2. Proportion of corresponding authors (CAs) in international collaborative papers in 2007 and 2011

<table>
<thead>
<tr>
<th>nation</th>
<th>2007</th>
<th></th>
<th></th>
<th>2011</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Co-paper</td>
<td>CAs</td>
<td>B/A</td>
<td>Co-paper</td>
<td>CAs</td>
<td>B/A</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td></td>
<td>A</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>20324</td>
<td>10975</td>
<td>54.0%</td>
<td>38398</td>
<td>21568</td>
<td>56.2%</td>
</tr>
</tbody>
</table>

Table 2 shows that the proportion of corresponding authors in China’s international collaboration is over 50%, and staying relatively stable in 2007 and 2011. Of course, the simply descriptive statistical data only give us some superficial impression, and the in-depth analysis need to use panel analysis tools to reveal the rules.

Which role did China take in the collaboration with the five other important nations? Table 3 shows that in collaboration with major scientific and technological powers, the proportion of corresponding authors is slightly lower than the average of China in 2007 and 2011, especially in collaboration with Germany and France. Therefore, in order to enhance China’s capacities in international scientific and technological collaboration, the proportion of corresponding authors needs to increase.

Table 3. Proportion of corresponding authors (CAs) between China and other 5 big nations in 2007 and 2011

<table>
<thead>
<tr>
<th>Co-nations</th>
<th>2007</th>
<th></th>
<th></th>
<th>2011</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Co-paper</td>
<td>CAs</td>
<td>B/A</td>
<td>Co-paper</td>
<td>CAs</td>
<td>B/A</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td></td>
<td>A</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>USA</td>
<td>7924</td>
<td>4032</td>
<td>50.9%</td>
<td>17065</td>
<td>9050</td>
<td>53.0%</td>
</tr>
<tr>
<td>JAPAN</td>
<td>2684</td>
<td>1363</td>
<td>50.8%</td>
<td>3915</td>
<td>2016</td>
<td>51.5%</td>
</tr>
<tr>
<td>UK</td>
<td>1992</td>
<td>1059</td>
<td>53.2%</td>
<td>3716</td>
<td>1985</td>
<td>53.4%</td>
</tr>
<tr>
<td>GERMANY</td>
<td>1683</td>
<td>793</td>
<td>47.1%</td>
<td>2861</td>
<td>1288</td>
<td>45.0%</td>
</tr>
<tr>
<td>FRANCE</td>
<td>1054</td>
<td>466</td>
<td>44.2%</td>
<td>1898</td>
<td>793</td>
<td>41.8%</td>
</tr>
</tbody>
</table>

How does the role differ according to discipline? Based on the SCI 22 subjects/disciplines we determine the proportion of corresponding authors in international collaborative papers for the years 2007 and 2011. The data shown in Table 4 reveal that in 2007, the following disciplines have a proportion of corresponding authors higher than the average (54.0%): agricultural sciences, chemistry, computer science, engineering, environment / ecology, geosciences, materials science, interdisciplinary, physics, plant and animal science. In the following disciplines the opposite is the case: clinical medicine, economics and business, immunology, molecular biology and genetics, neuroscience and behavior, social sciences (general). But in 2011, there are some changes. The following disciplines have a proportion of corresponding authors that is higher than average (56.2%): agricultural science, computer science, engineering, environment/ecology,
Some disciplines in which this proportion was higher than average in 2007 are now below average and vice versa. The collaborative counts of disciplines have doubled from 2007 to 2011; however, in terms of collaborative roles, in particular in terms of playing the leading role some disciplines, such as clinical medicine, immunology, molecular biology and genetics, and neuroscience & behavior, still show room for improvement.

Table 4. Proportions of corresponding authors (CAs) in collaborative disciplines in 2007 and 2011

<table>
<thead>
<tr>
<th>Disciplines</th>
<th>2007</th>
<th></th>
<th>2011</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Co-paper</td>
<td>CAs</td>
<td>B/A</td>
<td>Co-paper</td>
</tr>
<tr>
<td>Agricultural Sciences</td>
<td>406</td>
<td>228</td>
<td>56.2%</td>
<td>870</td>
</tr>
<tr>
<td>Biology &amp; Biochemistry</td>
<td>956</td>
<td>492</td>
<td>51.5%</td>
<td>1865</td>
</tr>
<tr>
<td>Chemistry</td>
<td>2586</td>
<td>1493</td>
<td>57.7%</td>
<td>4678</td>
</tr>
<tr>
<td>Clinical Medicine</td>
<td>2246</td>
<td>924</td>
<td>41.1%</td>
<td>5231</td>
</tr>
<tr>
<td>Computer Science</td>
<td>634</td>
<td>373</td>
<td>58.8%</td>
<td>1591</td>
</tr>
<tr>
<td>Economics &amp; Business</td>
<td>59</td>
<td>24</td>
<td>40.7%</td>
<td>127</td>
</tr>
<tr>
<td>Engineering</td>
<td>2452</td>
<td>1413</td>
<td>57.6%</td>
<td>5124</td>
</tr>
<tr>
<td>Environment/Ecology</td>
<td>786</td>
<td>444</td>
<td>56.5%</td>
<td>1570</td>
</tr>
<tr>
<td>Geosciences</td>
<td>1109</td>
<td>664</td>
<td>59.9%</td>
<td>1901</td>
</tr>
<tr>
<td>Immunology</td>
<td>190</td>
<td>76</td>
<td>40.0%</td>
<td>317</td>
</tr>
<tr>
<td>Materials Science</td>
<td>1431</td>
<td>872</td>
<td>60.9%</td>
<td>2656</td>
</tr>
<tr>
<td>Mathematics</td>
<td>934</td>
<td>476</td>
<td>51.0%</td>
<td>1438</td>
</tr>
<tr>
<td>Microbiology</td>
<td>347</td>
<td>170</td>
<td>49.0%</td>
<td>651</td>
</tr>
<tr>
<td>Molecular Biology &amp; Genetics</td>
<td>545</td>
<td>228</td>
<td>41.8%</td>
<td>1141</td>
</tr>
<tr>
<td>Multidisciplinary</td>
<td>244</td>
<td>133</td>
<td>54.5%</td>
<td>480</td>
</tr>
<tr>
<td>Neuroscience &amp; Behavior</td>
<td>390</td>
<td>157</td>
<td>40.3%</td>
<td>724</td>
</tr>
<tr>
<td>Pharmacology &amp; Toxicology</td>
<td>375</td>
<td>195</td>
<td>52.0%</td>
<td>621</td>
</tr>
<tr>
<td>Physics</td>
<td>3074</td>
<td>1760</td>
<td>57.3%</td>
<td>4607</td>
</tr>
<tr>
<td>Plant &amp; Animal Science</td>
<td>1031</td>
<td>588</td>
<td>57.0%</td>
<td>1684</td>
</tr>
<tr>
<td>Psychiatry/Psychology</td>
<td>80</td>
<td>39</td>
<td>48.8%</td>
<td>194</td>
</tr>
<tr>
<td>Social Sciences, general</td>
<td>125</td>
<td>55</td>
<td>44.0%</td>
<td>308</td>
</tr>
<tr>
<td>Space Science</td>
<td>323</td>
<td>171</td>
<td>52.9%</td>
<td>549</td>
</tr>
</tbody>
</table>

Overseas Chinese play the role of bridges in international collaboration

The connecting role of countrymen or former countrymen working in other countries has been studied in several publications (Webster, 2004; Jin et al, 2007; Jonkers & Tijssen, 2008). When it comes to Chinese scientists we will use the term overseas Chinese referring to those scientists who are of Chinese descent, but have joined other nationalities, as well as those Chinese which temporarily work and study in other nations. Among the 38,327 Chinese international collaborative papers in 2011, there
are 25,995 papers including overseas Chinese, about 67.8%.

The proportion of overseas Chinese collaborative papers in 22 disciplines (Table 5) shows that in some disciplines, such as physics (75.1%), interdisciplinary studies (74.2%), immunology (73.5%), materials science (73.5%), molecular biology and genetics (73.2%), and computer science (72.2%), the overseas Chinese are very active, often leading the collaboration. Even in some disciplines with a lower proportion of overseas Chinese collaborative papers, such as plant & animal science (50.4%). In sum, the overseas Chinese have a very strong role as a bridge in international collaboration with mainland China.

**Table 5. Proportion of overseas Chinese collaborative papers in 22 disciplines in 2011**

<table>
<thead>
<tr>
<th>Disciplines</th>
<th>2011 total numbers of collaborative papers</th>
<th>2011 numbers of Overseas Chinese papers</th>
<th>proportion of overseas Chinese papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agricultural Sciences</td>
<td>870</td>
<td>492</td>
<td>56.6%</td>
</tr>
<tr>
<td>Biology &amp; Biochemistry</td>
<td>1865</td>
<td>1335</td>
<td>71.6%</td>
</tr>
<tr>
<td>Chemistry</td>
<td>4678</td>
<td>3146</td>
<td>67.3%</td>
</tr>
<tr>
<td>Clinical Medicine</td>
<td>5231</td>
<td>3508</td>
<td>67.1%</td>
</tr>
<tr>
<td>Computer Science</td>
<td>1591</td>
<td>1148</td>
<td>72.2%</td>
</tr>
<tr>
<td>Economics &amp; Business</td>
<td>127</td>
<td>80</td>
<td>63.0%</td>
</tr>
<tr>
<td>Engineering</td>
<td>5124</td>
<td>3589</td>
<td>70.0%</td>
</tr>
<tr>
<td>Environment/Ecology</td>
<td>1570</td>
<td>961</td>
<td>61.2%</td>
</tr>
<tr>
<td>Geosciences</td>
<td>1901</td>
<td>1171</td>
<td>61.6%</td>
</tr>
<tr>
<td>Immunology</td>
<td>317</td>
<td>233</td>
<td>73.5%</td>
</tr>
<tr>
<td>Materials Science</td>
<td>2656</td>
<td>1953</td>
<td>73.5%</td>
</tr>
<tr>
<td>Mathematics</td>
<td>1438</td>
<td>867</td>
<td>60.3%</td>
</tr>
<tr>
<td>Microbiology</td>
<td>651</td>
<td>430</td>
<td>66.1%</td>
</tr>
<tr>
<td>Molecular Biology &amp; Genetics</td>
<td>1141</td>
<td>835</td>
<td>73.2%</td>
</tr>
<tr>
<td>Multidisciplinary</td>
<td>480</td>
<td>356</td>
<td>74.2%</td>
</tr>
<tr>
<td>Neuroscience &amp; Behavior</td>
<td>724</td>
<td>515</td>
<td>71.1%</td>
</tr>
<tr>
<td>Pharmacology &amp; Toxicology</td>
<td>621</td>
<td>442</td>
<td>71.2%</td>
</tr>
<tr>
<td>Physics</td>
<td>4607</td>
<td>3459</td>
<td>75.1%</td>
</tr>
<tr>
<td>Plant &amp; Animal Science</td>
<td>1684</td>
<td>851</td>
<td>50.5%</td>
</tr>
<tr>
<td>Psychiatry/Psychology</td>
<td>194</td>
<td>102</td>
<td>52.6%</td>
</tr>
<tr>
<td>Social Sciences, general</td>
<td>308</td>
<td>172</td>
<td>55.8%</td>
</tr>
<tr>
<td>Space Science</td>
<td>549</td>
<td>350</td>
<td>63.8%</td>
</tr>
</tbody>
</table>

**Conclusion and discussion**

This article analyzed the trends of international collaboration in science and technology using six leading nations (the United States, China, Japan, Germany, UK and France) as case studies. These data retrieved from the WoS show that the numbers of collaborative papers in five nations have grown much faster than the numbers of
non-collaborative ones. This means that the major scientific and technological nations focus on their own research capability while further promoting science and technology through international collaboration. The numbers of collaborative nations are continuously growing all over the world, but especially in China.

As a not-yet-fully-developed nation, China is playing an increasingly important role in international collaboration. Numbers of collaborative or non-collaborative papers have experienced a breakthrough growth. China’s international collaboration in science and technology has obvious disciplinary preferences, focusing mainly on traditional fundamental disciplines. Overseas Chinese play an extremely important role in mainland China’s international collaboration in science and technology, providing an optimal channel to overcome linguistic and cultural barriers. It is no surprise that China's S&T policies support and facilitate such collaborations.

Collaboration is an optimal strategy to promote science and technology

Why does the data show sharply increasing patterns of international collaboration in recent years?

In researchers’ level, collaboration is intrinsic needs in their careers. Everyone want to publish instead of perish, especially to publish in high level journal, such as SCIENCE or NATURE. Many data have verified the fact that partnering with collaborators anywhere can help them do the research, and even more importantly, get their papers accepted, especially partnering the prestigious professors. In earlier stage, the collaboration has taken place in colleagues, supervisor and students, even different institutes. With the support of IT, the collaboration gradually spread more remote distance, such as different states or nations. It can be called bottom-up pattern, which is driven by the researcher himself and is a self-organized process. The limitations of this pattern are stochastic and non-planned, but it is a critical factor to promote the sharp increment of collaboration, whether the local collaboration or international one.

International collaboration in science and technology besides being an effective way to boost productivity also promotes prosperity in the world. No nation, developing or developed, has enough resources to match its needs in science and technology. Therefore international collaboration has become an important aspect in scientific policies. We even claim that it is the optimal path to acquire needed resources. This is obvious in the case of Big Science projects such as CERN’s Large Hadron Collider (LHC).

Developed nations often have first-mover advantages in research and have enough resources to attract talented, excellent intellectuals from all over the world to participate in international collaboration. Due to their strength, they can be more selective with respect to the nature of their collaboration. In order to reach the research front of a discipline, developed nations have paid a lot of attention to increase their own scientific and technological capabilities, while promoting international collaboration and formulating appropriate international collaboration strategies.
Due to limitations in scientific and technological conditions it is impossible for developing or non-developed nations alike to carry out comprehensive research in all disciplines. Hence they have to formulate suitable strategies focusing on their best disciplines on the one hand, and promoting their other disciplines through international collaboration. Generally, developing nations often have an advantage in the number of human personnel, so that they have manpower to offer when collaborating with developed nations. Such collaborations may accelerate the development of science and technology in the world.

Therefore, international collaboration in science and technology is an optimal way to reallocate scientific and technological resources on a global scale. Every country sees international collaboration as a tool to develop its own R&D strength. This collaboration can be called top-down pattern, which is an optimal strategy driven by government policies. Not only the developing countries, such as China, but some developed ones, such as UK, have formulated some planning to promote international collaboration in science and technology. The impressive characteristics of this pattern are the target-oriented and the planned, and it can be used to interpret the preferred collaborative disciplines, institutions, and countries.

20 years ago or earlier, it is difficult to seek a collaborator in remote distance, especially different countries. It is the main thing that the advent of the internet has given more convenient condition to promote collaboration during the past two decades, especially the international cooperation. With the development of the Internet, scientific and technological collaboration in an international context has become easier. Gone are the days of hand written letters, difficult intercontinental telephone talks and an occasional long travel. Nowadays colleagues living on the other side of the globe are as close for discussions as a colleague working two stages higher in the building.

All of these, not only the needs of single researchers but the governmental policies and context changes of information techniques, together promote the international collaboration in science and technology. The bottom-up practices of collaborative needs in scientists have accelerated the increment of collaborative papers and settled the micro-foundation of collaboration, the top-down collaborative policies formulated by the government have facilitated the policies context, and the information techniques, particularly the internet, have eliminated the spatial barriers of communication and collaboration in long distances. The prosperities of international collaboration are due to the harmonious function of above factors.

Ethnic ties are a bridge to optimize collaborative channels for China

For developing nations, it is an important issue to seek an international collaborative partner in science and technology. Yet there exist many barriers to collaboration, such as culture, language, funds and policies. It is known that building on ethnic ties is an optimal channel to eliminate such barriers (Jin et al., 2007).

The ‘Overseas Chinese phenomenon’ (Jin et al., 2007) in China’s international
collaboration has developed from a deep social basis over many years. Chinese scientists leaving China, permanently or temporarily, are embedded in traditional Chinese culture, writing the same language, and having close social relations with China. These factors form a strong social basis binding overseas Chinese to a country with which they share many characteristics. This social basis is also where the opportunity lays for overseas Chinese to carry on international collaboration with China. From the foreign country’s perspective, the presence of a significant community of ethnic Chinese makes it easier to collaborate with scientists in mainland China. Indeed, Jonkers and Tijssen (2008) conclude that while countries may lose human capital when Chinese or other ethnic groups return to their homeland (a phenomenon referred to as “return brain drain”) they may also gain in terms of scientific linkages.

Furthermore, the government of mainland China has formulated several specific policies regarding overseas Chinese, such as *The Recruitment Program of Global Experts*, *Introducing Talent intellectuals Planning*, *Yangtze River Scholars Program*, *Thousand Talents program*, and so on. These programs attract outstanding overseas Chinese persuading them to return to their homeland.

Such ethnic ties are of a homophilic nature (McPherson et al., 2001), referring to the phenomenon that similar groups have more frequent contact than un-similar ones. On the one hand, we notice that overseas Chinese scientists establish emotional contact through family, relatives, teacher-student relationship and former collaboration, making future collaboration easier; and on the other hand, based on ethnic ties, this collaboration can enhance the prestige of participants in mainland China. In fact, this phenomenon may not only exist for Chinese, but also for other ethnic groups such as Indians and Koreans.

**References**


Multi-Scale H-Index: A New Measure to Assess the Scientific Impact of Scholars

Feng Ma¹ and Yuan Huang²

¹phoenixma@ahau.edu.cn, ²huy_921@163.com
¹Anhui Agricultural University, Hefei (China)
²HeFei University of Technology, Hefei (China)

Abstract
Quantitative evaluation on the individual’s scientific impact is an important research topic in the scientometrics and bibliometrics fields. As a pioneering milestone, the h-index measures the scientific impact of an individual’s research work by simultaneously considering the publication productivity and the article impact (citation count). However, the h-index ignores the additional citation count of those papers in the h-core publication list. To address this issue, lots of research efforts have been devoted and many alternatives are proposed but still suffer some limitations. In this paper, we extend the classic h-index and propose a novel multi-scale h-index to measure the scientific impact of scholars. In our approach, we first generate a multi-scale representation of a scholar’s paper citation record and calculate the h-index score for each scale. In this way, a multi-dimensional h-index score vector is obtained. To obtain a single number for quantitative evaluation and comparison, a final index score is calculated by summarizing those multi-scale h-index scores in an adaptive weighing manner. Our approach can be extended to consider all publication citations and subtly emphasizes those important papers with high citation count. To validate the proposed indicators, an experimental study is conducted on 82 scholars in the scientometrics field and comparison is made with several existing indices. Our proposed MH-index demonstrates to be more balanced and fine-grained for evaluating and comparing the scientific impact of individual scientists.

Conference Topic
Indicators

Introduction
In the scientific field, there is an important research problem, i.e., how to quantify the cumulative scientific significance and broad impact of an individual scholar. A proper overall quantitative index will provide a useful indicator to evaluate and compare different individuals competing for the same resource, such as universities faculty application or advancement and grant award, when scientific achievement is a critical assessment criterion (Hirsch 2005). To this end, generally, the publication and citation record of an individual is used to extract an overall indicator number. Early criteria commonly used include total number of papers, total number of citations, average citations per paper, number of significant papers, etc. Although intuitive, those criteria suffer from the imbalance between productivity and paper importance or the involvement of arbitrary parameters.

To avoid the disadvantage of the above criteria, Hirsch (2005) proposed the classic h-index as an easily computable indicator to characterize the scientific output of a researcher. As a pioneering work, the h-index measures the scientific impact of an individual by counting the maximal number of core papers with citation times over a threshold equal to the core paper number. Although simple and useful, the h-index suffers a nontrivial limitation that no additional credit is given to those highly cited articles, whereas every citation counts on scientific evaluation (Vanclay 2007, Rousseau et al. 2008, Schreiber 2008, Bornmann et al. 2008 and 2011). To address this weakness, many indicators are proposed, such as g-index (Egghe 2006), AR-index (Jin 2007), R-index (Jin et al. 2007), Tapered h-index (Anderson et al. 2008), Hg-index (Alonso et al. 2010), q-2 index (Cabrerizo et al. 2010), PR-index (Gao et al. 2016), ch-index (Kosmulski 2006), t-index (Tol 2009), pi-index (Vinkler 2009), w-index (Wu 2011).
To give credit to excess citation count, some other index schemes are proposed, such as e-index by Zhang (2009) and j-index by Todeschini (2011). Another limitation of h-index is that it relies on only those articles which have significant impact and high citation record, whereas every article contributes to research output and scientific impact of an individual. To consider the importance of all cited articles of an individual, several other indices are proposed, such as the multidimensional h-index by García Pérez (2009), mock h-index by Prathap (2010) central interval index by Dorta-González et al. (2011), two sided h-index by García Pérez (2012). To compare different index methods, comprehensive studies have been made by Bornmann et al. (2008 and 2011). Considering the difficulty to evaluate the scientific impact of an individual with a single indicator, some other researchers resort to the combination of two or multiple indicators (Martin 1996). Although such an idea seems interesting and promising, special concern shall be taken on the combination or aggregation scheme.

In this paper, we discuss the properties of h-index, e-index, multidimensional h-index, iteratively weighted h-index (iw(h)-index), EM-index, and propose a new measure named multi-scale h-index to evaluate the scientific impact of a scholar. The multi-scale h-index only considers those highly cited articles, but also counts those outside the h-core article set.

The rest of this paper is organized as follows. In “Related Work” section, we make a survey on some preliminary work including h-index, e-index, multidimensional h-index, iteratively weighted h-index, and EM-index, and discuss their advantage together with limitations. In “Multi-Scale H-Index” section, we elaborate our novel multi-scale h-index and its properties. After that, we conduct experimental study in “Experiments” section. Last but not least, we conclude this work in “Conclusion” section.

**Related Work**

This work is focused on the quantitative indicator of scientific impact based on an individual’s record of article publication and citation. In the following, several related indices are surveyed and analysed.

As the milestone index to measure an individual’s scientific impact, h-index (Hirsch, 2005) is defined based on the citation of an individual’s publication. The h-index of a scholar is $h$ if $h$ of his or her articles have at least $h$ citations each while the citation of each of the rest papers is no larger than $h$. Geometrically, given the citation curve of all papers in descending order, its intersection with the of the 45 degree line (the line $y = x$) gives $h$, as illustrated in Fig. 1(a).

![Figure 1](image-url)

**Figure 1.** Schematic illustration of (a) h-index, (b) multi-dimensional h-index, and (c) EM-index. The e-index is defined based on h-index with excess citation count considered, while the iteratively weighted h-index is a global representation of the multi-dimensional h-index.
The h-index only considers those high-cited articles but ignores those with less-cited ones. To address this problem, the multi-dimensional h-index is proposed. Given the citation record of a scientist, his or her multi-dimensional h-index \((h_1, h_2, h_3, \cdots, h_K)\) is generated as follows. All the involved \(N\) articles are sorted by the citation count in descending order. Then, \(h_1\) is obtained as the traditional h-index. After that, those top \(h_1\) most-cited articles are removed and we compute \(h_2\) as the h-index of the remaining \((N-h_1)\) articles. Similarity, \(h_{k+1}\) is obtained by computing the h-index on the bottom \((N-h_1-h_2-\cdots-h_k)\) least cited articles, which are obtained by checking the intersection of the line \(y = x - \sum_{i=1}^{k} h_i\) with the article citation curve, as illustrated in Fig. 1 (b). Such process is iterated until all articles with citation count no less than 1 are visited. Since multi-dimensional h-index is a vector and inconvenient for comparison, a global number is defined by the iteratively weighted h-index (Todeschini and Baccini 2016) as follows,

\[
iw(h) = \sum_{k=1}^{K} \frac{h_k}{k}.
\] (1)

To discriminate those highly-cited articles, some variants of h-index are proposed. Zhang et al. (2009) proposed e-index to consider the importance of the excess citation count over h-index. Specifically, given the sorted article citations \((c_1, c_2, \cdots)\) in descending order of a scientist, the e-index is defined as square root of sum of the excess citation count of \(h\)-core articles, as follows,

\[
e = \sqrt{\sum_{k=1}^{h} (c_k - h)}.
\] (2)

Another variant is EM-index (Bihari and Tripathi 2017), which adopts the concept of multi-dimensional h-index. Concretely speaking, for an individual with article citations \((c_1, c_2, \cdots)\) ranked in descending order, EM-index first generates a multi-dimensional index vector \((h_1, h_2, h_3, \cdots, h_K)\). The first component \(h_1\) is exactly the h-index value. The \((k+1)\)th component \(h_{k+1}\) is obtained by computing the h-index for the excess citation counts of the \(h\)-core papers in the \(k\)-th round, which are obtained by checking the intersection of the line \(y = x - \sum_{i=1}^{k} h_i\) with the article citation curve, as illustrated in Fig. 1 (c). Finally, the EM-index is defined as the square root of the sum over the all components in the index vector.

The above index schemes only consider the highly cited articles, but fail to explore all articles with at least one citation. To make a comprehensive evaluation, both the top and the tail citation counts should be taken into consideration. Towards this goal, García Pérez (2012) proposed a two-sided h-index. Bihari et al. (2017) extended their EM-index to EM'-index by considering the citation counts of all articles after subtracting the h-index from the h-core articles.

**Multi-Scale II-Index**

To evaluate an individual’s scientific impact based on the publication and citation records, the general principle beneath the existing schemes, such as h-index, e-index, iterative weighted h-index and EM-index, is to correlate the publication count with the citation count under some constraints. Following such a paradigm, an intuitive idea is to define a measure as: the index value of an individual is \(m\), if at most the \(m\) articles have citation counts over \(n\) while the citation counts of the remaining articles are all no larger than \(n\). A function \(f(\cdot)\) is defined to correlate \(m\) and \(n\) as follows,

\[
m = f(n).
\] (3)

In h-index, the correlation function is an identity function \(f(x) = x\), which is inherited by many other variants. Although some a function is simple and demonstrates effective in many cases, it is not flexible to well capture the fine-grained characteristics of the citation curve. Based on such a motivation, we propose to define the correlation function as follows,

\[
y = f_k(x) = 2^k x,
\] (4)
where $k$ is an integer. It is notable that, when $k$ equals 0, the above correlation function degenerates to that of the classic h-index. By checking the intersection of the line function $y = f_k(x)$ with the citation count curve, we can obtain the corresponding index component $h_k$. By varying the value of $k$, a multi-dimensional index vector $(\cdots, h_{-2}, h_{-1}, h_0, h_1, h_2, \cdots)$ is generated, as illustrated in Fig. 2. Each component $h_k > 0$ and $h_0$ is equal to the classic h-index value. With a larger $k$, the corresponding component $h_k$ considers articles with much higher citations.

Another perspective to interpret our scheme is the multi-scale representation of the descending citation curve. In the above discussion, we fix the citation curve $y = g(x)$ but change the correlations functions with varying values of $k$. Alternatively, we can also fix all the correlation function $y = f_k(x)$ as the identity function $f(x) = x$, but transform the citation curve $y = g(x)$ into multi-scale representations: $y = g_k(x) = 2^{-k}g(x)$. As a result, in the $k$-th scale, after the transformation of the citation curve, we can calculate the component $h_k$ as the classic h-index on the transformed citation curve. In this sense, we call our scheme as multi-scale (MS) h-index. We define two versions of multi-scale h-index. As the basic multi-scale h-index (MH-index), we impose that $k$ is non-negative, as illustrated in Fig. 2(a). Based on that, we extend it by keeping the full component vector with all possible $k$ and call it the extended multi-scale h-index abbreviated as MH(ext)-index, as illustrated in Fig. 2(b).

To generate a global number for convenient of evaluation and comparison, the multi-scale (MS) h-index is defined as follows,

$$MS = \sqrt{\sum_k w_k h_k},$$

where $w_k$ is a weighting constant to discriminate the component $h_k$ is different scale. Considering the above multi-scale formulation, $w_k$ is defined as $\alpha^k$, where $\alpha > 1$, to give more credit to more highly cited article. In our experiments, we set $\alpha = \sqrt{2}$.

![Figure 2. Schematic illustration of our proposed (a) multi-scale h-index and (b) the extended multi-scale h-index](image)

<table>
<thead>
<tr>
<th>Citation Counts</th>
<th>h-index</th>
<th>iw(h)</th>
<th>EM-index</th>
<th>MH-index</th>
<th>MH(ext)-index</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁ = (10, 8, 5, 5, 4, 0, 0)</td>
<td>4</td>
<td>4.50</td>
<td>3.16</td>
<td>3.70</td>
<td>5.07</td>
</tr>
<tr>
<td>C₂ = (100, 8, 5, 5, 4, 0, 0)</td>
<td>4</td>
<td>4.50</td>
<td>10</td>
<td>5.60</td>
<td>6.59</td>
</tr>
<tr>
<td>C₃ = (10, 10, 10, 10, 10, 10, 10)</td>
<td>8</td>
<td>8</td>
<td>3.16</td>
<td>4.68</td>
<td>5.25</td>
</tr>
<tr>
<td>C₄ = (10, 2, 1, 1, 1, 1, 0)</td>
<td>2</td>
<td>3.28</td>
<td>3.16</td>
<td>2.87</td>
<td>4.35</td>
</tr>
</tbody>
</table>
To illustrate the advantage of our proposed multi-scale h-index, we compare with h-index, iterative-weighted h-index (iw(h)-index) and EM-index on four synthetic examples in Table 1. Intuitively, the author with citation count C1 has less scientific impact than the author with C2. However, both h-index and iterative-weighted h-index get the same index values for both authors and fail to discriminate them. In contrast, our multi-scale h-index distinguish C1 from C2 with notable difference. On the other hand, it is obvious that the scientific impact of C1 is less than that of C3 but greater than that of C4. However, the EM-index assigns them with identical index values while our multi-scale h-index clearly differentiate them with the correct impact rank.

Experiments
To evaluate our proposed multi-scale h-index scheme, we collect the citation data of 82 scholars (Bihari and Tripathi 2017) in the scientometrics and bibliometrics fields from Google Scholar. The total citation counts of those scholars range from 390 to 128,909, while the publication article number varies from 27 to 598. This dataset contains scholars with both high productivity and high citation count, scholars with high productivity and less citation count, and scholars with average productivity and average citation count. The total publication and citation together with the h-index of the 82 authors are illustrated in Fig. 3.

![Figure 3. The total citation number, paper number and h-index of the studied 82 authors. The authors are sorted based on their MH-index scores in descending order. The data collection date from Google Scholar is April 26, 2017.](image_url)

We compare our proposed MH-index with h-index, iterative-weighted h-index, and EM-index. The results of each index scheme on the 82 scholars are listed in Table 2. In Table 3, we show the rank of the authors based on the index scores calculated with each methods. Our proposed MH-index provides a more fine-grained representation on the publication and citation records. Distinguished from h-index, multi-dimensional h-index and iteratively weighted h-index, our approach gives credit to those highly cited articles. On the other hand, our MH-index is different from EM-index in that, we calculate the significance index factor in a multi-scale paradigm and assigns more weight to the significance factor from articles of larger citations. The above properties make our MH-index produce more consistent and better ranking results on the scholar’s scientific impact. More case study results are discussed below.

For instance, for the scholars Linton C. Freeman and Stan Wasserman, their h-index values are 49 and 45, respectively, while their iw(h)-index values are 66.47 and 61.41, respectively. As a result, they are ranked to no. 23 and no. 29 by h-index, and no. 33 and no. 38 by iw(h)-index. In fact, compared with the top-ranking authors, the publication article number of Linton C. Freeman (109 articles) and Stan Wasserman (106 articles) is relatively few. However, both of them have quite a few very highly cited articles with remarkable total citation (35424 and...
Their top 1 citation counts are 27,680 and 11,329, respectively. Those extremely high citations are ignored by h-index and iteratively weighted h-index. In contrast, our MH-index makes a full consideration of this fact and rank them to no. 2 and no. 4, respectively. Another interesting example is Heidi Winklhofer, whose h-index is 22 but EM-index is 63.46. As a result, the scholar is ranked to as high as no. 9 by EM-index, which is unreasonable considering the relatively less publication article number (31) and total citation (6270). That is due to the fact that EM-index extremely favours the case like Heidi Winklhofer, whose top 1 citation is 4027 but the following one takes a cliff-like drop to 433. Our MH-index well address such cases and rank Heidi Winklhofer to no. 38.

Table 2. Comparative result of h-index, iteratively weighted h-index, EM-index, MH-index and the extended MH-index

<table>
<thead>
<tr>
<th>Author Name</th>
<th>h-index</th>
<th>iw(h)-index</th>
<th>EM-index</th>
<th>MH-index</th>
<th>MH(ext)-index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mark Newman</td>
<td>91</td>
<td>125.16</td>
<td>125.10</td>
<td>44.97</td>
<td>48.42</td>
</tr>
<tr>
<td>Loet Leydesdorff</td>
<td>87</td>
<td>154.57</td>
<td>77.96</td>
<td>29.96</td>
<td>37.19</td>
</tr>
<tr>
<td>C. Lee Giles</td>
<td>85</td>
<td>148.48</td>
<td>49.03</td>
<td>28.17</td>
<td>35.65</td>
</tr>
<tr>
<td>Richard S.J. Tol</td>
<td>87</td>
<td>158.20</td>
<td>32.71</td>
<td>26.44</td>
<td>35.01</td>
</tr>
<tr>
<td>Linton C. Freeman</td>
<td>49</td>
<td>66.47</td>
<td>106.39</td>
<td>30.60</td>
<td>33.18</td>
</tr>
<tr>
<td>Hirsch J.E.</td>
<td>63</td>
<td>104.98</td>
<td>80.99</td>
<td>27.00</td>
<td>32.16</td>
</tr>
<tr>
<td>Stan Wasserman</td>
<td>45</td>
<td>61.41</td>
<td>166.32</td>
<td>29.47</td>
<td>32.01</td>
</tr>
<tr>
<td>Steve Lawrence</td>
<td>62</td>
<td>84.62</td>
<td>49.08</td>
<td>28.41</td>
<td>31.99</td>
</tr>
<tr>
<td>Matthew O. Jackson</td>
<td>69</td>
<td>97.77</td>
<td>59.69</td>
<td>27.54</td>
<td>31.93</td>
</tr>
<tr>
<td>Adamantios Diamantopoulos</td>
<td>68</td>
<td>98.70</td>
<td>63.46</td>
<td>26.84</td>
<td>31.40</td>
</tr>
<tr>
<td>Enrique Herrera-Viedma</td>
<td>70</td>
<td>110.00</td>
<td>36.81</td>
<td>25.72</td>
<td>31.38</td>
</tr>
<tr>
<td>Santo Fortunato</td>
<td>42</td>
<td>61.48</td>
<td>73.16</td>
<td>29.80</td>
<td>29.80</td>
</tr>
<tr>
<td>Wolfgang Glänzel</td>
<td>63</td>
<td>107.45</td>
<td>24.74</td>
<td>21.92</td>
<td>28.49</td>
</tr>
<tr>
<td>Claes Wohlin</td>
<td>52</td>
<td>89.23</td>
<td>64.13</td>
<td>22.66</td>
<td>28.02</td>
</tr>
<tr>
<td>Gerhard Woeginger</td>
<td>59</td>
<td>107.55</td>
<td>41.06</td>
<td>20.69</td>
<td>27.96</td>
</tr>
<tr>
<td>Albert Zomaya</td>
<td>53</td>
<td>108.90</td>
<td>20.83</td>
<td>18.95</td>
<td>27.87</td>
</tr>
<tr>
<td>Michael Jennions</td>
<td>53</td>
<td>85.64</td>
<td>39.28</td>
<td>22.70</td>
<td>27.62</td>
</tr>
<tr>
<td>Ben R Martin</td>
<td>43</td>
<td>69.83</td>
<td>46.55</td>
<td>22.85</td>
<td>26.89</td>
</tr>
<tr>
<td>Yannis Manolopoulos</td>
<td>50</td>
<td>91.17</td>
<td>45.14</td>
<td>20.62</td>
<td>26.87</td>
</tr>
<tr>
<td>Guang-Hong Yang</td>
<td>52</td>
<td>108.31</td>
<td>22.47</td>
<td>17.95</td>
<td>26.78</td>
</tr>
<tr>
<td>Roberto Todeschini</td>
<td>45</td>
<td>75.68</td>
<td>58.61</td>
<td>21.96</td>
<td>26.61</td>
</tr>
<tr>
<td>András Schubert</td>
<td>51</td>
<td>78.85</td>
<td>49.28</td>
<td>21.86</td>
<td>26.43</td>
</tr>
<tr>
<td>Anthony F.J. van Raan</td>
<td>56</td>
<td>85.77</td>
<td>25.98</td>
<td>21.30</td>
<td>26.32</td>
</tr>
<tr>
<td>Ronald Rousseau</td>
<td>47</td>
<td>93.46</td>
<td>30.46</td>
<td>19.03</td>
<td>26.22</td>
</tr>
<tr>
<td>Andrew D. Jackson</td>
<td>60</td>
<td>93.24</td>
<td>28.02</td>
<td>20.56</td>
<td>26.21</td>
</tr>
<tr>
<td>Anne-Wil Harzing</td>
<td>51</td>
<td>70.77</td>
<td>24.74</td>
<td>22.49</td>
<td>26.13</td>
</tr>
<tr>
<td>Blaise Cronin</td>
<td>49</td>
<td>88.30</td>
<td>30.36</td>
<td>19.22</td>
<td>25.60</td>
</tr>
<tr>
<td>Josep Domingo-Ferrer</td>
<td>48</td>
<td>90.11</td>
<td>23.17</td>
<td>18.75</td>
<td>25.45</td>
</tr>
<tr>
<td>Lutz Bornmann</td>
<td>47</td>
<td>89.20</td>
<td>25.90</td>
<td>18.62</td>
<td>25.39</td>
</tr>
<tr>
<td>Henk F. Moed</td>
<td>51</td>
<td>75.52</td>
<td>36.70</td>
<td>20.76</td>
<td>25.24</td>
</tr>
<tr>
<td>Vicenc Torra</td>
<td>41</td>
<td>76.11</td>
<td>29.27</td>
<td>18.74</td>
<td>24.63</td>
</tr>
<tr>
<td>Mark Fine</td>
<td>47</td>
<td>78.05</td>
<td>27.77</td>
<td>18.51</td>
<td>23.91</td>
</tr>
<tr>
<td>Johan Bollen</td>
<td>34</td>
<td>51.97</td>
<td>50.20</td>
<td>20.48</td>
<td>23.65</td>
</tr>
<tr>
<td>Name</td>
<td>Age</td>
<td>First_name</td>
<td>Second_name</td>
<td>First</td>
<td>Last</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>-----</td>
<td>------------</td>
<td>-------------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>Herbert Van de Sompel</td>
<td>38</td>
<td>59.37</td>
<td>25.53</td>
<td>19.08</td>
<td>23.10</td>
</tr>
<tr>
<td>Judit Bar-Ilan</td>
<td>41</td>
<td>63.98</td>
<td>23.19</td>
<td>17.87</td>
<td>22.38</td>
</tr>
<tr>
<td>Michael S. Rosenberg</td>
<td>34</td>
<td>42.86</td>
<td>34.45</td>
<td>20.05</td>
<td>22.29</td>
</tr>
<tr>
<td>Çağan Hakkı Şekercioğlu</td>
<td>36</td>
<td>53.16</td>
<td>23.45</td>
<td>18.96</td>
<td>22.29</td>
</tr>
<tr>
<td>Marek Kosmulski</td>
<td>34</td>
<td>58.78</td>
<td>24.64</td>
<td>17.78</td>
<td>22.26</td>
</tr>
<tr>
<td>Daniel HD</td>
<td>36</td>
<td>61.77</td>
<td>25.90</td>
<td>17.28</td>
<td>21.92</td>
</tr>
<tr>
<td>Serge Galam</td>
<td>35</td>
<td>60.57</td>
<td>19.13</td>
<td>17.25</td>
<td>21.88</td>
</tr>
<tr>
<td>Weiguo (Patrick) Fan</td>
<td>39</td>
<td>64.89</td>
<td>18.68</td>
<td>16.81</td>
<td>21.66</td>
</tr>
<tr>
<td>Kène Henkens</td>
<td>39</td>
<td>67.51</td>
<td>14.97</td>
<td>15.56</td>
<td>21.10</td>
</tr>
<tr>
<td>Roger Brumback</td>
<td>37</td>
<td>70.53</td>
<td>12.88</td>
<td>14.10</td>
<td>20.94</td>
</tr>
<tr>
<td>Jayant Vaidya</td>
<td>31</td>
<td>51.50</td>
<td>23.79</td>
<td>17.01</td>
<td>20.94</td>
</tr>
<tr>
<td>Benny Lautrup</td>
<td>36</td>
<td>54.41</td>
<td>21.38</td>
<td>16.78</td>
<td>20.67</td>
</tr>
<tr>
<td>Aric Hagberg</td>
<td>28</td>
<td>41.21</td>
<td>33.56</td>
<td>16.85</td>
<td>19.87</td>
</tr>
<tr>
<td>Maria Bordons</td>
<td>33</td>
<td>51.97</td>
<td>18.17</td>
<td>15.93</td>
<td>19.87</td>
</tr>
<tr>
<td>Nees Jan van Eck</td>
<td>31</td>
<td>45.14</td>
<td>23.56</td>
<td>16.56</td>
<td>19.84</td>
</tr>
<tr>
<td>Christoph Bartneck</td>
<td>30</td>
<td>51.62</td>
<td>21.05</td>
<td>14.89</td>
<td>19.41</td>
</tr>
<tr>
<td>Heidi Winklofer</td>
<td>21</td>
<td>25.25</td>
<td>63.46</td>
<td>17.92</td>
<td>19.37</td>
</tr>
<tr>
<td>Dimitrios Katsaros</td>
<td>27</td>
<td>42.19</td>
<td>27.91</td>
<td>15.95</td>
<td>19.20</td>
</tr>
<tr>
<td>Carlos Pecharroman</td>
<td>31</td>
<td>49.88</td>
<td>17.46</td>
<td>15.08</td>
<td>19.11</td>
</tr>
<tr>
<td>Fiorenzo Franceschini</td>
<td>31</td>
<td>54.87</td>
<td>15.10</td>
<td>13.89</td>
<td>19.00</td>
</tr>
<tr>
<td>Sergio Alonso</td>
<td>24</td>
<td>34.10</td>
<td>19.39</td>
<td>16.49</td>
<td>18.92</td>
</tr>
<tr>
<td>Peter Jacso</td>
<td>26</td>
<td>48.94</td>
<td>20.83</td>
<td>14.17</td>
<td>18.89</td>
</tr>
<tr>
<td>Berwin Turlach</td>
<td>21</td>
<td>35.35</td>
<td>27.13</td>
<td>16.05</td>
<td>18.87</td>
</tr>
<tr>
<td>Gangan Prathap</td>
<td>29</td>
<td>56.53</td>
<td>14.46</td>
<td>12.95</td>
<td>18.85</td>
</tr>
<tr>
<td>Paul Wouters</td>
<td>28</td>
<td>44.28</td>
<td>17.03</td>
<td>15.14</td>
<td>18.69</td>
</tr>
<tr>
<td>John Irvine</td>
<td>26</td>
<td>37.26</td>
<td>25.59</td>
<td>15.65</td>
<td>18.34</td>
</tr>
<tr>
<td>Hendrik P. van Dalen</td>
<td>28</td>
<td>50.22</td>
<td>12.37</td>
<td>13.29</td>
<td>18.15</td>
</tr>
<tr>
<td>Ludo Waltman</td>
<td>26</td>
<td>34.32</td>
<td>23.56</td>
<td>15.73</td>
<td>18.08</td>
</tr>
<tr>
<td>Sune Lehmann</td>
<td>17</td>
<td>24.20</td>
<td>35.11</td>
<td>15.66</td>
<td>17.46</td>
</tr>
<tr>
<td>Duncan Lindsey</td>
<td>25</td>
<td>36.50</td>
<td>21.79</td>
<td>14.41</td>
<td>17.37</td>
</tr>
<tr>
<td>Ruediger Mutz</td>
<td>23</td>
<td>35.50</td>
<td>19.52</td>
<td>13.95</td>
<td>17.00</td>
</tr>
<tr>
<td>Francisco Javier Cabrerozo</td>
<td>19</td>
<td>29.72</td>
<td>19.39</td>
<td>14.57</td>
<td>16.97</td>
</tr>
<tr>
<td>Jörn Altmann</td>
<td>23</td>
<td>43.54</td>
<td>13.78</td>
<td>11.55</td>
<td>16.53</td>
</tr>
<tr>
<td>Morten Schmidt</td>
<td>20</td>
<td>33.28</td>
<td>18.60</td>
<td>13.24</td>
<td>16.40</td>
</tr>
<tr>
<td>Rodrigo Costas</td>
<td>22</td>
<td>33.54</td>
<td>17.46</td>
<td>13.24</td>
<td>16.25</td>
</tr>
<tr>
<td>Miguel A. García-Pérez</td>
<td>22</td>
<td>41.40</td>
<td>14.73</td>
<td>10.80</td>
<td>15.56</td>
</tr>
<tr>
<td>Clint D. Kelly</td>
<td>18</td>
<td>27.85</td>
<td>17.35</td>
<td>12.72</td>
<td>15.27</td>
</tr>
<tr>
<td>Raj Kumar Pan</td>
<td>21</td>
<td>27.45</td>
<td>19.49</td>
<td>12.53</td>
<td>14.93</td>
</tr>
<tr>
<td>Birger Larsen Pan</td>
<td>21</td>
<td>38.92</td>
<td>8.12</td>
<td>9.76</td>
<td>14.57</td>
</tr>
<tr>
<td>Domenico A. Maisano</td>
<td>19</td>
<td>32.85</td>
<td>15.10</td>
<td>10.36</td>
<td>14.21</td>
</tr>
<tr>
<td>Yu-Hsin Liu</td>
<td>16</td>
<td>20.45</td>
<td>17.55</td>
<td>12.09</td>
<td>13.86</td>
</tr>
<tr>
<td>Luca Mastrogiacomilo</td>
<td>16</td>
<td>28.48</td>
<td>9.75</td>
<td>9.50</td>
<td>13.24</td>
</tr>
<tr>
<td>Alireza Abbasi</td>
<td>14</td>
<td>20.03</td>
<td>13.78</td>
<td>10.76</td>
<td>12.88</td>
</tr>
<tr>
<td>Nils T. Hagen</td>
<td>16</td>
<td>21.97</td>
<td>10.72</td>
<td>10.46</td>
<td>12.81</td>
</tr>
<tr>
<td>Ash Mohammad Abbas</td>
<td>12</td>
<td>20.49</td>
<td>8.60</td>
<td>7.97</td>
<td>10.83</td>
</tr>
</tbody>
</table>
Table 3. Comparative result of the author rank based on h-index, iteratively weighted h-index, EM-index, MH-index and the extended MH-index

<table>
<thead>
<tr>
<th>Author Name</th>
<th>Rank by h-index</th>
<th>Rank by iw(h)-index</th>
<th>Rank by EM-index</th>
<th>Rank by MH-index</th>
<th>Rank by MH(ext)-index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mark Newman</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Loet Leydesdorff</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>C. Lee Giles</td>
<td>4</td>
<td>3</td>
<td>15</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Richard S.J. Tol</td>
<td>3</td>
<td>1</td>
<td>25</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>Linton C. Freeman</td>
<td>23</td>
<td>33</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Hirsch J.E.</td>
<td>8</td>
<td>10</td>
<td>4</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>Stan Wasserman</td>
<td>29</td>
<td>38</td>
<td>1</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Steve Lawrence</td>
<td>10</td>
<td>23</td>
<td>14</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Matthew O. Jackson</td>
<td>6</td>
<td>13</td>
<td>10</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Adamantios Diamantopoulos</td>
<td>7</td>
<td>12</td>
<td>8</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Enrique Herrera-Viedma</td>
<td>5</td>
<td>5</td>
<td>20</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>Santo Fortunato</td>
<td>32</td>
<td>37</td>
<td>6</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>Wolfgang Glänzel</td>
<td>9</td>
<td>9</td>
<td>38</td>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>Claes Wohlin</td>
<td>17</td>
<td>18</td>
<td>7</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>Gerhard Woeginger</td>
<td>12</td>
<td>8</td>
<td>18</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>Albert Zomaya</td>
<td>15</td>
<td>6</td>
<td>52</td>
<td>32</td>
<td>16</td>
</tr>
<tr>
<td>Michael Jennions</td>
<td>16</td>
<td>22</td>
<td>19</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>Ben R Martin</td>
<td>31</td>
<td>31</td>
<td>16</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>Yannis Manolopoulos</td>
<td>22</td>
<td>16</td>
<td>17</td>
<td>23</td>
<td>19</td>
</tr>
<tr>
<td>Guang-Hong Yang</td>
<td>18</td>
<td>7</td>
<td>47</td>
<td>37</td>
<td>20</td>
</tr>
<tr>
<td>Roberto Todeschini</td>
<td>30</td>
<td>27</td>
<td>11</td>
<td>17</td>
<td>21</td>
</tr>
<tr>
<td>András Schubert</td>
<td>19</td>
<td>24</td>
<td>13</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>Mauno Vihinen</td>
<td>13</td>
<td>11</td>
<td>49</td>
<td>27</td>
<td>23</td>
</tr>
<tr>
<td>Anthony F.J. van Raan</td>
<td>14</td>
<td>21</td>
<td>33</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>Ronald Rousseau</td>
<td>26</td>
<td>14</td>
<td>26</td>
<td>30</td>
<td>25</td>
</tr>
<tr>
<td>Andrew D. Jackson</td>
<td>11</td>
<td>15</td>
<td>29</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>Anne-Wil Harzing</td>
<td>20</td>
<td>29</td>
<td>39</td>
<td>16</td>
<td>27</td>
</tr>
<tr>
<td>Blaise Cronin</td>
<td>24</td>
<td>20</td>
<td>27</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>Josep Domingo-Ferrer</td>
<td>25</td>
<td>17</td>
<td>46</td>
<td>33</td>
<td>29</td>
</tr>
<tr>
<td>Lutz Bornmann</td>
<td>27</td>
<td>19</td>
<td>34</td>
<td>35</td>
<td>30</td>
</tr>
<tr>
<td>Henk F. Moed</td>
<td>21</td>
<td>28</td>
<td>21</td>
<td>21</td>
<td>31</td>
</tr>
<tr>
<td>Vicenç Torra</td>
<td>33</td>
<td>26</td>
<td>28</td>
<td>34</td>
<td>32</td>
</tr>
<tr>
<td>Mark Fine</td>
<td>28</td>
<td>25</td>
<td>31</td>
<td>36</td>
<td>33</td>
</tr>
<tr>
<td>Johan Bollen</td>
<td>43</td>
<td>46</td>
<td>12</td>
<td>25</td>
<td>34</td>
</tr>
<tr>
<td>Herbert Van de Sompel</td>
<td>37</td>
<td>40</td>
<td>37</td>
<td>29</td>
<td>35</td>
</tr>
<tr>
<td>Judit Bar-Ilan</td>
<td>34</td>
<td>35</td>
<td>45</td>
<td>39</td>
<td>36</td>
</tr>
<tr>
<td>Michael S. Rosenberg</td>
<td>44</td>
<td>56</td>
<td>23</td>
<td>26</td>
<td>37</td>
</tr>
<tr>
<td>Çağan Hakkı Şekerçioğlu</td>
<td>39</td>
<td>45</td>
<td>44</td>
<td>31</td>
<td>38</td>
</tr>
<tr>
<td>Marek Kosmulski</td>
<td>45</td>
<td>41</td>
<td>40</td>
<td>40</td>
<td>39</td>
</tr>
<tr>
<td>Name</td>
<td>Page</td>
<td>First</td>
<td>Last</td>
<td>Total</td>
<td>Page</td>
</tr>
<tr>
<td>-----------------</td>
<td>------</td>
<td>-------</td>
<td>------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>Daniel HD</td>
<td>40</td>
<td>36</td>
<td>35</td>
<td>41</td>
<td>40</td>
</tr>
<tr>
<td>Serge Galam</td>
<td>42</td>
<td>39</td>
<td>58</td>
<td>42</td>
<td>41</td>
</tr>
<tr>
<td>Weiguo (Patrick) Fan</td>
<td>35</td>
<td>34</td>
<td>59</td>
<td>45</td>
<td>42</td>
</tr>
<tr>
<td>Kène Henkens</td>
<td>36</td>
<td>32</td>
<td>69</td>
<td>55</td>
<td>43</td>
</tr>
<tr>
<td>Roger Brumback</td>
<td>38</td>
<td>30</td>
<td>74</td>
<td>62</td>
<td>44</td>
</tr>
<tr>
<td>Jayant Vaidya</td>
<td>47</td>
<td>49</td>
<td>41</td>
<td>43</td>
<td>45</td>
</tr>
<tr>
<td>Benny Lautrup</td>
<td>41</td>
<td>44</td>
<td>50</td>
<td>46</td>
<td>46</td>
</tr>
<tr>
<td>Aric Hagberg</td>
<td>53</td>
<td>59</td>
<td>24</td>
<td>44</td>
<td>47</td>
</tr>
<tr>
<td>Maria Bordons</td>
<td>46</td>
<td>47</td>
<td>61</td>
<td>51</td>
<td>48</td>
</tr>
<tr>
<td>Nees Jan van Eck</td>
<td>48</td>
<td>53</td>
<td>42</td>
<td>47</td>
<td>49</td>
</tr>
<tr>
<td>Christoph Bartneck</td>
<td>51</td>
<td>48</td>
<td>51</td>
<td>58</td>
<td>50</td>
</tr>
<tr>
<td>Heidi Winklhofer</td>
<td>66</td>
<td>75</td>
<td>9</td>
<td>38</td>
<td>51</td>
</tr>
<tr>
<td>Dimitrios Katsaros</td>
<td>56</td>
<td>57</td>
<td>30</td>
<td>50</td>
<td>52</td>
</tr>
<tr>
<td>Carlos Pecharroman</td>
<td>49</td>
<td>51</td>
<td>63</td>
<td>57</td>
<td>53</td>
</tr>
<tr>
<td>Fiorenzo Franceschini</td>
<td>50</td>
<td>43</td>
<td>67</td>
<td>64</td>
<td>54</td>
</tr>
<tr>
<td>Sergio Alonso</td>
<td>61</td>
<td>66</td>
<td>56</td>
<td>48</td>
<td>55</td>
</tr>
<tr>
<td>Peter Jacso</td>
<td>57</td>
<td>52</td>
<td>53</td>
<td>61</td>
<td>56</td>
</tr>
<tr>
<td>Berwin Turlach</td>
<td>67</td>
<td>64</td>
<td>32</td>
<td>49</td>
<td>57</td>
</tr>
<tr>
<td>Gangan Prathap</td>
<td>52</td>
<td>42</td>
<td>71</td>
<td>68</td>
<td>58</td>
</tr>
<tr>
<td>Paul Wouters</td>
<td>54</td>
<td>54</td>
<td>66</td>
<td>56</td>
<td>59</td>
</tr>
<tr>
<td>John Irvine</td>
<td>58</td>
<td>61</td>
<td>36</td>
<td>54</td>
<td>60</td>
</tr>
<tr>
<td>Hendrik P. van Dalen</td>
<td>55</td>
<td>50</td>
<td>75</td>
<td>65</td>
<td>61</td>
</tr>
<tr>
<td>Ludo Waltman</td>
<td>59</td>
<td>65</td>
<td>43</td>
<td>52</td>
<td>62</td>
</tr>
<tr>
<td>Sune Lehmann</td>
<td>75</td>
<td>76</td>
<td>22</td>
<td>53</td>
<td>63</td>
</tr>
<tr>
<td>Duncan Lindsey</td>
<td>60</td>
<td>62</td>
<td>48</td>
<td>60</td>
<td>64</td>
</tr>
<tr>
<td>Ruediger Mutz</td>
<td>62</td>
<td>63</td>
<td>54</td>
<td>63</td>
<td>65</td>
</tr>
<tr>
<td>Francisco Javier Cabrerizo</td>
<td>71</td>
<td>70</td>
<td>57</td>
<td>59</td>
<td>66</td>
</tr>
<tr>
<td>Jörn Altmann</td>
<td>63</td>
<td>55</td>
<td>72</td>
<td>72</td>
<td>67</td>
</tr>
<tr>
<td>Morten Schmidt</td>
<td>70</td>
<td>68</td>
<td>60</td>
<td>66</td>
<td>68</td>
</tr>
<tr>
<td>Rodrigo Costas</td>
<td>64</td>
<td>67</td>
<td>64</td>
<td>67</td>
<td>69</td>
</tr>
<tr>
<td>Miguel A. García-Pérez</td>
<td>65</td>
<td>58</td>
<td>70</td>
<td>73</td>
<td>70</td>
</tr>
<tr>
<td>Clint D. Kelly</td>
<td>74</td>
<td>72</td>
<td>65</td>
<td>69</td>
<td>71</td>
</tr>
<tr>
<td>Raj Kumar Pan</td>
<td>68</td>
<td>73</td>
<td>55</td>
<td>70</td>
<td>72</td>
</tr>
<tr>
<td>Birger Larsen</td>
<td>69</td>
<td>60</td>
<td>81</td>
<td>78</td>
<td>73</td>
</tr>
<tr>
<td>Domenico A. Maisano</td>
<td>72</td>
<td>69</td>
<td>68</td>
<td>77</td>
<td>74</td>
</tr>
<tr>
<td>Yu-Hsin Liu</td>
<td>76</td>
<td>79</td>
<td>62</td>
<td>71</td>
<td>75</td>
</tr>
<tr>
<td>Luca Mastrogiacomo</td>
<td>77</td>
<td>71</td>
<td>78</td>
<td>79</td>
<td>76</td>
</tr>
<tr>
<td>András Teles</td>
<td>73</td>
<td>74</td>
<td>77</td>
<td>76</td>
<td>77</td>
</tr>
<tr>
<td>Alirez Abbasi</td>
<td>79</td>
<td>80</td>
<td>73</td>
<td>74</td>
<td>78</td>
</tr>
<tr>
<td>Nils T. Hagen</td>
<td>78</td>
<td>77</td>
<td>76</td>
<td>75</td>
<td>79</td>
</tr>
<tr>
<td>Ash Mohammad Abbas</td>
<td>80</td>
<td>78</td>
<td>79</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>Raf Guns</td>
<td>81</td>
<td>81</td>
<td>80</td>
<td>81</td>
<td>81</td>
</tr>
<tr>
<td>Fred Y. Ye</td>
<td>82</td>
<td>82</td>
<td>82</td>
<td>82</td>
<td>82</td>
</tr>
</tbody>
</table>
Conclusion
In this paper, we propose a novel multi-scale h-index to measure the scientific impact of individual scholars. We mathematically formulate the index scheme from the perspective of correlation function to collaboratively investigate the publication and citation record. With a multi-scale paradigm, a multi-dimension index factor is generated, which is further summarized by adaptive weighting into a global index number. The proposed MH-index makes a fine-grained measure and gives adaptive credit based on the significance level of the publication article citation count. We evaluate our index scheme with comprehensive study on the real scholars and quantitatively demonstrate the advantage of the proposed MH-index over the comparison indices.

References
García Pérez, M. A. (2012). An extension of the h index that covers the tail and the top of the citation curve and allows ranking researchers with similar h. *Journal of Informetrics*, 6(4), 689–699.


A principled methodology for comparing relatedness measures for clustering publications

Ludo Waltman¹  Kevin W. Boyack²  Giovanni Colavizza³  Nees Jan van Eck¹

¹{ludo.waltman, nees.jan.van.eck}@cwts.leidenuniv.nl
Centre for Science and Technology Studies, Leiden University (The Netherlands)

²kboyack@mapofscience.com
SciTech Strategies, Inc., Albuquerque, NM (USA)

³ giovanni.colavizza@epfl.ch
Digital Humanities Laboratory, École Polytechnique Fédérale de Lausanne (Switzerland)

Abstract
There are many different relatedness measures, based for instance on citation relations or textual similarity, that can be used to cluster scientific publications. We propose a principled methodology for comparing relatedness measures and for evaluating the accuracy of clustering solutions resulting from these measures. Our empirical results, based on publications in the field of condensed matter physics, suggest that bibliographic coupling relations yield more accurate clustering solutions than direct citation relations and co-citation relations. The so-called extended direct citation approach is competitive with bibliographic coupling in terms of the accuracy of the resulting clustering solutions. Our results also suggest that BM25 yields more accurate clustering solutions than other text-based relatedness measures.

Conference topic
Citation and co-citation analysis; methods and techniques

Introduction
Clustering of scientific publications is an important bibliometric problem. Bibliometricians have employed many different clustering techniques (e.g., Gläser, Scharnhorst, & Glänzel, 2017). Moreover, they have also used various different relatedness measures to cluster publications. These relatedness measures are typically based on either citation relations (e.g., direct citation relations, bibliographic coupling relations, or co-citation relations) or textual similarity, or sometimes a combination of the two.

Which relatedness measure yields the most accurate clustering of publications? Two perspectives can be taken on this question. One perspective is that there is no absolute notion of accuracy (e.g., Gläser et al., 2017). Following this perspective, each relatedness measure yields clustering solutions that are accurate in their own right, and it is not meaningful to ask which clustering solutions are the overall most accurate ones. Different citation-based and text-based relatedness measures for instance each emphasize different aspects of the way in which publications relate to each other, and the corresponding clustering solutions each provide a legitimate viewpoint on the organization of the scientific literature. The other perspective is that for some purposes it is useful, and perhaps even necessary, to assume the existence of an absolute notion of accuracy. When this perspective is taken, it is possible, at least in principle, to say that some relatedness measures yield more accurate clustering solutions than others.

We believe that both perspectives are useful. From a purely conceptual point of view, the first perspective is probably the more satisfactory one. However, from a more applied point of
view, the second perspective is highly important. In many practical applications, users expect to be provided with a single clustering of publications. Users typically have some intuitive idea of accuracy and, based on this idea of accuracy, they expect the clustering provided to them to be as accurate as possible. In this paper, we take the applied viewpoint and we therefore focus on the second perspective.

Identifying the relatedness measure that yields the most accurate clustering of publications is challenging because of the lack of a ground truth. There is no perfect clustering of publications that can be used to evaluate the accuracy of different clustering solutions. As an alternative, the accuracy of clustering solutions can be evaluated by domain experts who assess the quality of different clustering solutions in a specific scientific domain (e.g., Šubelj, Van Eck, & Waltman, 2016). The difficulty of this approach is that it is hard to find a sufficiently large number of experts who are willing to spend a considerable amount of time to make a detailed assessment of the quality of different clustering solutions. Moreover, the knowledge of experts will often be restricted to relatively small domains, and it will be unclear to what extent the conclusions drawn by experts generalize to other domains.

In this paper, we take a large-scale data-driven approach to compare different relatedness measures based on which publications can be clustered. The basic idea is to cluster publications based on a number of different relatedness measures and to use another more or less independent relatedness measure to evaluate the accuracy of the clustering solutions. This approach has already been used extensively in a series of papers by Kevin Boyack, Dick Klavans, and colleagues. They compared different citation-based relatedness measures (Boyack & Klavans, 2010; Klavans & Boyack, 2017), including relatedness measures that take advantage of full-text data (Boyack, Small, & Klavans, 2013), as well as different text-based relatedness measures (Boyack et al., 2011). To evaluate the accuracy of clustering solutions, they used grant data and textual similarity (Boyack & Klavans, 2010; Boyack et al., 2011, 2013) and more recently also the reference lists of ‘authoritative’ publications, defined as publications with at least 100 references (Klavans & Boyack, 2017).

Our aim in this paper is to introduce a more principled methodology for performing analyses similar to the ones mentioned above. We restrict ourselves to the use of one specific clustering technique, namely the technique introduced in the bibliometric literature by Waltman and Van Eck (2012), but we allow the use of any measure of the relatedness of publications. For two relatedness measures $A$ and $B$, our proposed methodology offers a principled way to evaluate the accuracy of clustering solutions obtained using the two measures, where a third relatedness measure $C$ is used as the evaluation criterion.

**Methodology**

To introduce our methodology for evaluating the accuracy of clustering solutions obtained using different relatedness measures, we first discuss the quality function that we use to cluster publications. We then explain how we evaluate the accuracy of a clustering solution, and we analyze the consistency of our evaluation framework. Finally, we discuss the importance of using an independent evaluation criterion.

**Quality function for clustering publications**

Consider a set of $N$ publications. Let $r_{ij}^X \geq 0$ denote the relatedness of publications $i$ and $j$ (with $i = 1,\ldots,N$ and $j = 1,\ldots,N$) based on relatedness measure $X$, and let $c_i^X \in \{1,2,\ldots\}$ denote the cluster to which publication $i$ belongs when publications have been clustered based on relatedness measure $X$.

Publications are assigned to clusters by maximizing a quality function. We focus on the quality function of Waltman and Van Eck (2012). This quality function is given by
\[ Q = \sum_{i,j} I(c_i^X = c_j^X)(r_{ij}^X - \gamma), \]  

(1)

where \( I(c_i^X = c_j^X) \) equals 1 if \( c_i^X = c_j^X \) and 0 otherwise and where \( \gamma \geq 0 \) denotes a so-called resolution parameter. The higher the value of this parameter, the larger the number of clusters that will be obtained. Hence, the resolution parameter \( \gamma \) determines the granularity of the clustering.

The quality function in (1) can also be written as

\[ Q = \sum_{i,j} I(c_i^X = c_j^X)r_{ij}^X - \gamma \sum (s_k^X)^2, \]  

(2)

where \( s_k^X \) denotes the number of publications belonging to cluster \( k \), that is,

\[ s_k^X = \sum_i I(c_i^X = k). \]  

(3)

We also refer to \( s_k^X \) as the size of cluster \( k \).

In the network science literature, the above quality function was proposed by Traag, Van Dooren, and Nesterov (2011), who referred to it as the constant Potts model. The quality function is closely related to the well-known modularity function introduced by Newman and Girvan (2004) and Newman (2004), but it has the important advantage that it does not suffer from the so-called resolution limit problem (Fortunato & Barthélémy, 2007). Waltman and Van Eck (2012) introduced the above quality function in the bibliometric literature.

**Evaluating the accuracy of a clustering solution**

Suppose that we have three relatedness measures \( A, B, \) and \( C \), and suppose also that we have used relatedness measures \( A \) and \( B \) to cluster a set of publications. Furthermore, suppose that we want to use relatedness measure \( C \) to evaluate the accuracy of the clustering solutions obtained using relatedness measures \( A \) and \( B \). Let \( A^{X|C} \) denote the accuracy of a clustering solution obtained using relatedness measure \( X \) (with \( X = A \) or \( X = B \)), where the accuracy is evaluated using relatedness measure \( C \). We define \( A^{X|C} \) as

\[ A^{X|C} = \frac{1}{N} \sum_{i,j} I(c_i^X = c_j^X)r_{ij}^C. \]  

(4)

The clustering solution obtained using relatedness measure \( A \) is considered to be more accurate than the clustering solution obtained using relatedness measure \( B \) if \( A^{A|C} > A^{B|C} \), and the other way around.

The above approach for evaluating the accuracy of a clustering solution favors less granular solutions over more granular ones. Of all possible clustering solutions, the least granular solution is the one in which all publications belong to the same cluster. According to (4), this least granular clustering solution always has the highest possible accuracy. There can be no other clustering solution with a higher accuracy. In order to perform meaningful comparisons, (4) should be used only for comparing clustering solutions that have the same granularity.

How do we determine whether two clustering solutions have the same granularity? We could require that both clustering solutions have been obtained using the same value for the resolution parameter \( \gamma \). Alternatively, we could require that both clustering solutions consist of the same number of clusters. We do not take either of these approaches. Instead, we require that the sum of the squared cluster sizes is the same for two clustering solutions. In other words, two clustering solutions obtained using relatedness measures \( A \) and \( B \) have the same granularity if
\[ \sum_k (s_k^A)^2 = \sum_l (s_l^B)^2. \]  

(5)

If (5) is satisfied, (4) can be used to compare in an unbiased way the clustering solutions obtained using relatedness measures \( A \) and \( B \). On the other hand, if (5) is not satisfied, a comparison based on (4) will be biased in favor of the less granular clustering solution. In practice, obtaining two clustering solutions that satisfy (5) typically will not be trivial. For each of the two clustering solutions, it may require a significant amount of trial and error with different values of the resolution parameter \( \gamma \). In the end, it may turn out that (5) can be satisfied only approximately, not exactly. We will get back to this issue in the discussion of our empirical results.

Consistency of the evaluation framework

The choice of the accuracy measure defined in (4) and the granularity condition presented in (5) may seem quite arbitrary. However, provided that we use the quality function defined in (1), this choice has an important justification. When the accuracy of clustering solutions is evaluated using some relatedness measure \( X \), our choice of the accuracy measure in (4) and the granularity condition in (5) guarantees that of all possible clustering solutions of a certain granularity the solution obtained using relatedness measure \( X \) will be the most accurate one. In other words, it is guaranteed that \( A^X|X \geq A^Y|X \) for any relatedness measure \( Y \). This is a fundamental consistency property that we believe should be satisfied by any sound framework for evaluating the accuracy of clustering solutions obtained using different relatedness measures.

Suppose for instance that we have three clustering solutions, all of the same granularity, one solution obtained based on direct citation relations between publications, another one obtained based on bibliographic coupling relations, and a third one obtained based on co-citation relations. Suppose also that the accuracy of the clustering solutions is evaluated based on direct citation relations. It would then be a rather odd outcome if the clustering solution obtained based on bibliographic coupling or co-citation relations turned out to be more accurate than the solution obtained based on direct citation relations. In our evaluation framework, it is guaranteed that there can be no such inconsistent outcomes. When the accuracy of clustering solutions is evaluated based on direct citation relations, the clustering solution obtained based on direct citation relations will always be the most accurate one.

Independent evaluation criterion

As already mentioned, the approach that we take in this paper is to cluster publications based on a number of different relatedness measures and to use another more or less independent relatedness measure to evaluate the accuracy of the clustering solutions. More specifically, we use a text-based relatedness measure to evaluate the accuracy of different clustering solutions obtained using citation-based relatedness measures, and the other way around, we use a citation-based relatedness measure to evaluate the accuracy of different clustering solutions obtained using text-based relatedness measures.

Importantly, we do not evaluate citation-based clustering solutions using a citation-based relatedness measure or text-based clustering solutions using a text-based relatedness measure. Such evaluations do not provide much information, because the relatedness measure used for evaluation is not sufficiently independent from the relatedness measures being evaluated. For instance, when direct citation relations are used to evaluate the accuracy of different clustering solutions obtained using citation-based relatedness measures, the clustering solution obtained based on direct citation relations can be expected to be the most accurate one. We do not learn much from this. The evaluation simply shows that the clustering
solution obtained based on direct citation relations is best aligned with an evaluation criterion based on direct citation relations, which of course is not surprising at all. This illustrates the importance of using an independent evaluation criterion. The more the relatedness measure used for evaluation can be considered to be independent of the relatedness measures being evaluated, the more informative the evaluation will be.

**Relatedness measures**

We now discuss the relatedness measures that we consider in this paper. We first discuss relatedness measures based on citation relations, followed by relatedness measures based on textual similarity. We also discuss the idea of the so-called top M relatedness approach combined with the idea of normalized relatedness measures.

**Citation-based relatedness measures**

Below we discuss a number of citation-based approaches for determining the pairwise relatedness for a set of $N$ publications. We use $c_{ij}$ to indicate whether publication $i$ cites publication $j$ ($c_{ij} = 1$) or not ($c_{ij} = 0$). The relatedness of publications $i$ and $j$ based on direct citation relations is given by

$$r_{ij}^{DC} = c_{ij} + c_{ji}. \quad (6)$$

Hence, $r_{ij}^{DC} = 1$ if either publication $i$ cites publication $j$ or the other way around and $r_{ij}^{DC} = 0$ if neither publication cites the other.

The relatedness of publications $i$ and $j$ based on bibliographic coupling relations equals the number of common references in the two publications. This can be written as

$$r_{ij}^{BC} = \sum_k c_{ik} c_{jk}, \quad (7)$$

where the summation extends over all publications in the database that we use, not only over the $N$ publications for which we aim to determine their pairwise relatedness.

As is well known, co-citation can be seen as the opposite of bibliographic coupling. The relatedness of publications $i$ and $j$ based on co-citation relations equals the number of publications in which publications $i$ and $j$ are both cited. In mathematical terms,

$$r_{ij}^{CC} = \sum_k c_{ki} c_{kj}, \quad (8)$$

where the summation again extends over all publications in the database that we use.

The above approaches for determining the relatedness of publications may also be combined. This results in

$$r_{ij}^{DC-BC-CC} = \alpha r_{ij}^{DC} + r_{ij}^{BC} + r_{ij}^{CC}, \quad (9)$$

where $\alpha$ denotes a parameter that determines the weight of direct citation relations relative to bibliographic coupling and co-citation relations. A direct citation relation may be considered a stronger signal of the relatedness of two publications than a bibliographic coupling or co-citation relation (Waltman & Van Eck, 2012), and therefore one may want to give more weight to a direct citation relation than to the two other types of relations. This can be achieved by setting $\alpha$ to a value above 1.

In addition to the above citation-based approaches for determining the relatedness of publications, we also consider a so-called extended direct citation approach. Like the ordinary
direct citation approach, the extended direct citation approach takes into account only direct citation relations between publications. However, direct citation relations are considered not just within the set of $N$ focal publications but within an extended set of publications. In addition to the $N$ focal publications, the extended set of publications includes all publications in our database that have a direct citation relation with at least two focal publications. The technical details of the extended direct citation approach are somewhat complex. Because of space limitations, we do not discuss these details. We note that an approach similar to our extended direct citation approach was also used by Boyack and Klavans (2014) and Klavans and Boyack (2017).

**Text-based relatedness measures**

We consider two text-based approaches for determining the relatedness of publications. We use $o_{il}$ to denote the number of occurrences of term $l$ in publication $i$. To count the number of occurrences of a term in a publication, only the title and abstract of the publication are considered, not the full text. Part-of-speech tagging is applied to the title and abstract of the publication to identify nouns and adjectives. A term is defined as a sequence of nouns and adjectives, with the last word in the sequence being a noun. No distinction is made between singular and plural nouns, so *neural network* and *neural networks* are regarded as the same term. Furthermore, shorter terms embedded in longer terms are not counted. For instance, if a publication contains the term *artificial neural network*, this is counted as an occurrence of *artificial neural network* but not as an occurrence of *neural network* or *network*. Finally, no stop word list is used, so there are no terms that are excluded from being counted. A straightforward text-based measure of the relatedness of publications $i$ and $j$ is given by

$$ r_{ij}^{CT} = \sum_l \frac{o_{il}o_{jl}}{\left(\sum_k o_{kl}\right)^{\beta}}. \quad (10) $$

We refer to this as relatedness based on common terms. The denominator in (10) aims to reduce the influence of frequently occurring terms. The parameter $\beta$ in the denominator determines the extent to which the influence of these terms is reduced. If $\beta = 0$, no reduction in the influence of frequently occurring terms takes place. On the other hand, if $\beta = 1$, the influence of frequently occurring terms is strongly reduced, following a so-called fractional counting approach (Perianes-Rodriguez, Waltman, & Van Eck, 2016).

Boyack et al. (2011) identified BM25 as one of the most accurate text-based relatedness measures for clustering publications. We therefore also include BM25 in our analysis. BM25 originates from the field of information retrieval, where it is used to determine the relevance of a document for a search query (Sparck Jones, Walker, & Robertson, 2000a, 2000b). Following Boyack et al. (2011), we use BM25 as a text-based measure of the relatedness of publications. The BM25 relatedness measure is defined as

$$ r_{ij}^{BM25} = \sum_l I(o_{il} > 0) \text{IDF}_l \frac{o_{jl}(k_1+1)}{o_{jl+k_1}\left(1-b+bd\right)}, \quad (11) $$

where $I(o_{il} > 0)$ equals 1 if $o_{il} > 0$ and 0 otherwise and where $d_j$ and $\bar{d}$ denote, respectively, the length of publication $j$ and the average length of all $N$ publications. We define the length of a publication as the total number of occurrences of terms in the publication. This results in

$$ d_i = \sum_l o_{il} \quad \text{and} \quad \bar{d} = \frac{1}{N} \sum_i d_i. \quad (12) $$

IDF$_l$ in (11) denotes the inverse document frequency of term $l$, which we define as

$$ \text{IDF}_l = \log \left(\frac{N}{\# \text{of documents containing } l}\right), $$

where $\# \text{of documents containing } l$ is the number of publications containing term $l$. The constant $k_1$ is set to 1.2 and $b$ is set to 0.75. The length of a publication is defined as the total number of occurrences of terms in the publication.
\[
\text{IDF}_l = \log \frac{N - n_l + 0.5}{n_l + 0.5},
\]

where \( n_l \) denotes the number of publications in which term \( l \) occurs, that is,

\[
n_l = \sum_i I(o_{il} > 0).
\]

The BM25 relatedness measure in (11) depends on the parameters \( k_1 \) and \( b \). Following Boyack et al. (2011), we set these parameters to values of 2 and 0.75, respectively. We note that, unlike all other relatedness measures that we consider in this paper, the BM25 relatedness measure is not symmetrical. In other words, \( r_{ij}^{BM25} \) does not need to be equal to \( r_{ji}^{BM25} \).

**Top M relatedness approach and normalized relatedness measures**

Our interest focuses on large-scale clustering analyses that may involve hundreds of thousands or even millions of publications. These analyses impose significant challenges in terms of computing time and memory requirements. In particular, it may not be feasible in these analyses to store all non-zero relatedness values in a computer’s main memory.

To deal with this problem, we use the top \( M \) relatedness approach. This approach is quite similar to the idea of similarity filtering typically used by Kevin Boyack and Dick Klavans (e.g., Boyack & Klavans, 2010; Boyack et al., 2011). In the top \( M \) relatedness approach, only the top \( M \) strongest relations per publication are kept. The remaining relations are discarded.

We use \( \hat{r}_{ij}^X \) to denote the relatedness of publications \( i \) and \( j \) based on relatedness measure \( X \) after discarding relations that are not in the top \( M \) per publication. This means that \( \hat{r}_{ij}^X = r_{ij}^X \) if publication \( j \) is among the \( M \) publications that are most strongly related to publication \( i \) and that \( \hat{r}_{ij}^X = 0 \) otherwise. Relatedness of a publication with itself is ignored. Hence, \( \hat{r}_{ij}^X = 0 \) if \( i = j \). We note that even if \( r_{ij}^X \) is symmetrical, \( \hat{r}_{ij}^X \) does not need to be symmetrical.

In the analyses presented in this paper, we use a value of 20 for \( M \). We apply the top \( M \) relatedness approach to all our relatedness measures except for the measures based on direct citation relations. As pointed out by Waltman and Van Eck (2012), the use of direct citation relations has the advantage of requiring only a relatively limited amount of computer memory, and therefore there is no need to use the top \( M \) relatedness approach when working with direct citation relations. Applying the top \( M \) relatedness approach in the case of direct citation relations would also be problematic because all relations are equally strong, making it difficult to decide which relations to keep and which ones to discard. Hence, in the case of direct citation relations, we simply have \( \hat{r}_{ij}^{DC} = r_{ij}^{DC} \) for all publications \( i \) and \( j \).

We also normalize all relatedness measures. The normalized relatedness of publication \( i \) with publication \( j \) equals the relatedness of the publications divided by the total relatedness of publication \( i \) with all other publications. Hence, the relatedness of publication \( i \) with publication \( j \) based on relatedness measure \( X \) is given by

\[
\hat{r}_{ij}^X = \frac{\hat{r}_{ij}^X}{\sum_k \hat{r}_{ik}^X}.
\]

This normalization was also used by Waltman and Van Eck (2012). The idea of the normalization is that relatedness values of publications in different fields of science should be of the same order of magnitude, so that clusters in different fields will be of similar size. Without the normalization, citation-based relatedness values for instance can be expected to
be much higher in the life sciences than in the social sciences. In a clustering analysis that involves both publications in the life sciences and publications in the social sciences, this would result in life science clusters being systematically larger than social science clusters. The normalization in (15) can be used to correct for these kinds of differences between fields. All results presented in the next section are based on normalized relatedness measures.

Results
We start the discussion of the results of our analyses by explaining the data collection and the way in which publications were clustered. We then introduce the idea of granularity-accuracy plots. Next, we present a comparison of different citation-based relatedness measures that can be used to cluster publications. This is followed by a comparison of different text-based relatedness measures.

Data collection and clustering of publications
Data was collected from the Web of Science database. We used the in-house version of the Web of Science database available at the Centre for Science and Technology Studies at Leiden University. This version of the database includes the Science Citation Index Expanded, the Social Sciences Citation Index, and the Arts & Humanities Citation Index. Like in our earlier work (e.g., Klavans & Boyack, 2017; Waltman & Van Eck, 2012), our final interest is in clustering all publications available in the database that we use, without restricting ourselves to certain fields of science. However, to keep the analyses presented in this paper manageable, we did restrict ourselves to one specific field. We selected all publications of the document types article and review that appeared in the period 2007–2016 in journals belonging to the Web of Science subject category Physics, condensed matter. The number of selected publications is 272,935.

The citation-based and text-based relatedness measures discussed above were calculated for the selected publications. Two comments need to be made. First, in determining bibliographic coupling relations between publications, only common references to publications indexed in our database were considered. Common references to non-indexed publications (e.g., books, conference proceedings publications, and PhD theses) were not taken into account. Non-indexed publications were not considered in the extended direct citation approach either. Second, our database included a limited number of publications from 2017. These publications were not used in determining co-citation relations between publications. They also were not considered in the extended direct citation approach.

The selected publications were clustered based on each of our relatedness measures. Clustering was performed by maximizing the quality function presented in (1). To maximize the quality function, we used an iterative variant (Waltman & Van Eck, 2013) of the well-known Louvain algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). Five iterations of the algorithm were performed. In addition, to speed up the algorithm, we employed ideas similar to the pruning idea of Ozaki, Tezuka, and Inaba (2016) and the prioritization idea of Bae, Halperin, West, Rosvall, and Howe (2017). Our algorithm usually will not be able to find the global maximum of the quality function, but it can be expected to get close to the global maximum.

Different levels of granularity were considered. For each relatedness measure, we obtained ten clustering solutions, each of them for a different value of the resolution parameter $\gamma$. The following values of $\gamma$ were used: 0.00001, 0.00002, 0.00005, 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, and 0.01.
Granularity-accuracy plots

A difficulty of our evaluation framework is the requirement that the clustering solutions being compared have exactly the same granularity. This requirement, which is formalized in the condition in (5), is hard to meet in practice. Clustering solutions obtained using different relatedness measures but the same value of the resolution parameter $\gamma$ will approximately satisfy (5), but the condition normally will not be satisfied exactly.

To deal with this problem, we propose a graphical approach based on the idea of granularity-accuracy (GA) plots. Using a GA plot, relatedness measures can be compared despite differences in granularity between clustering solutions. The horizontal axis in a GA plot represents the granularity of a clustering solution. We define the granularity of a clustering solution obtained using relatedness measure $X$ as

$$\frac{N}{\sum_k (s_{X,k})^2}.$$  \hspace{1cm} (16)

We note that two clustering solutions that have the same granularity according to (16) indeed satisfy the condition in (5). The vertical axis in a GA plot represents the accuracy of a clustering solution as defined in (4). Clustering solutions are plotted in a GA plot based on their granularity and accuracy. Lines are drawn between clustering solutions obtained using the same relatedness measure but different values of the resolution parameter $\gamma$. In this paper, we use a logarithmic scale for both the horizontal and the vertical axis in a GA plot.

In the interpretation of a GA plot, one should be aware that for any relatedness measure an increase in granularity will always cause a decrease in accuracy. This is a mathematical necessity in our evaluation framework, and therefore it is not something that one should be concerned about. A GA plot can be interpreted by comparing the accuracy of different relatedness measures at a specific level of granularity. As explained above, clustering solutions obtained using different relatedness measures normally do not have exactly the same granularity. However, in a GA plot, lines are drawn between different clustering solutions obtained using the same relatedness measure, providing interpolations between these solutions. Based on such interpolations, the accuracy of different relatedness measures can be compared at a specific level of granularity. These comparisons can be performed at all levels of granularity. Sometimes different levels of granularity will yield inconsistent results, with for instance relatedness measure $A$ outperforming relatedness measure $B$ at one level of granularity and the opposite outcome at another level of granularity. In other cases, consistent results will be obtained at all levels of granularity. For instance, relatedness measure $C$ may consistently outperform relatedness measure $D$, regardless of the level of granularity.

In the next two subsections, GA plots will be used to compare different citation-based and text-based relatedness measures.

Comparison of citation-based relatedness measures

Figure 1 presents a GA plot for comparing the DC, BC, CC, DC-BC-CC, and EDC citation-based relatedness measures. In the case of the DC-BC-CC relatedness measure, two values of the parameter $\alpha$ are considered, $\alpha = 1$ and $\alpha = 5$. The BM25 text-based relatedness measure is used as the evaluation criterion. Results obtained when this relatedness measure is used to cluster publications are also included in the GA plot. These results provide an upper bound for the results that can be obtained using the citation-based relatedness measures. (Recall that the highest possible accuracy is obtained when publications are clustered based on the same relatedness measure that is also used as the evaluation criterion.)

To interpret the GA plot in Figure 1, it is important to have some understanding of the meaning of the different levels of granularity. When the granularity is around 0.001, a
clustering solution typically has about 500 significant clusters, defined as clusters that include at least ten publications. A granularity around 0.01 corresponds with a typical number of about 4,500 significant clusters. Finally, when the granularity is around 0.1, there typically are about 8,000 significant clusters. We emphasize that these statistics have a limited accuracy. The number of significant clusters depends quite strongly on the relatedness measure that is used.

Figure 1 shows that, when BM25 is used as the evaluation criterion, CC has the worst performance of all citation-based relatedness measures. This is not surprising. Uncited publications have no co-citation relations with other publications and therefore cannot be properly clustered. This is an important explanation of the bad performance of CC. The bad performance of CC is also in line with recent results of Klavans and Boyack (2017). DC outperforms CC but is outperformed by all other citation-based relatedness measures. The somewhat disappointing performance of DC is an important finding, especially given the increasing popularity of DC in recent years. BC, DC-BC-CC, and EDC all perform about equally well. DC-BC-CC and EDC seem to slightly outperform BC, but the difference is tiny. Likewise, there is hardly any difference between the parameter values $\alpha = 1$ and $\alpha = 5$ for DC-BC-CC. Our finding that BC and EDC perform about equally well differs from results of Klavans and Boyack, who found that an approach similar to EDC significantly outperforms BC. Our results are based on a more principled evaluation framework and a different evaluation criterion than the results of Klavans and Boyack, which explains why our findings are different from theirs.

Comparison of text-based relatedness measures

Figure 2 presents a GA plot for comparing the BM25 and CT text-based relatedness measures. In the case of the CT relatedness measure, three values of the parameter $\beta$ are considered, $\beta = 0.0$, $\beta = 0.5$, and $\beta = 1.0$. The DC-BC-CC citation-based relatedness measure (with $\alpha = 1$) is used as the evaluation criterion. Results obtained when this relatedness measure is used to cluster publications are also included in the GA plot. These results provide an upper bound for the results that can be obtained using the text-based relatedness measures.
The results presented in Figure 2 are quite clear. Using DC-BC-CC as the evaluation criterion, BM25 outperforms CT, regardless of the value of the parameter $\beta$. The good performance of BM25 is in agreement with the results of Boyack et al. (2011). By far the worst performance is obtained when CT is used with the parameter value $\beta = 0.0$. This confirms the importance of reducing the influence of frequently occurring terms.

Figure 2. GA plot for comparing text-based relatedness measures. The DC-BC-CC citation-based relatedness measure (with $\alpha = 1$) is used as the evaluation criterion.

Conclusions
The problem of clustering scientific publications involves significant conceptual and methodological challenges. In this paper, we have introduced a principled methodology for evaluating the accuracy of clustering solutions obtained using different relatedness measures. Our methodology can be applied to evaluate the accuracy of clustering solutions obtained using two relatedness measures $A$ and $B$, where a third relatedness measure $C$ is used as the evaluation criterion. Preferably, relatedness measure $C$ should be as independent as possible from relatedness measures $A$ and $B$. Relatedness measures $A$ and $B$ for instance may be citation-based relatedness measures, and relatedness measure $C$ may be a text-based relatedness measure (or the other way around).

The empirical results that we have presented are based on a large-scale analysis of publications in the field of condensed matter physics indexed in the Web of Science database. We have used our proposed methodology, complemented with a graphical approach based on so-called GA plots, to compare different citation-based relatedness measures that can be used to cluster publications. Using the BM25 text-based relatedness measure as the evaluation criterion, we have found that co-citation relations and direct citation relations yield less accurate clustering solutions than a number of other citation-based relatedness measures. Bibliographic coupling relations, possibly combined with direct citation relations and co-citation relations, can be used to obtain much more accurate clustering solutions. The accuracy of these clustering solutions is similar to the accuracy of clustering solutions obtained using the so-called extended direct citation approach. The other way around, we have compared different text-based relatedness measures using a citation-based relatedness measure (obtained by combining direct citation relations, bibliographic coupling relations, and co-citation relations) as the evaluation criterion. BM25 has turned out to yield more accurate clustering solutions than the other text-based relatedness measures that we have studied.
Acknowledgement

We would like to thank Dick Klavans and Vincent Traag for their helpful comments on our work.

References


A new diversity indicator based on a similarity treatment expansion of the Gini coefficient

Nicolas Bérubé¹, Vincent Larivière²

¹ nicolas.berube.3@umontreal.ca
École de bibliothéconomie et des sciences de l’information, Université de Montréal, Montréal (Canada)

² vincent.lariviere@umontreal.ca
École de bibliothéconomie et des sciences de l’information, Université de Montréal, Montréal (Canada) and Observatoire des sciences et des Technologies (OST), Centre Interuniversitaire de recherche sur la science et la technologie (CIRST), Montréal (Canada)

Abstract
Diversity indicators have recently taken into account the similarity between classes. Among those new indicators, the Leinster-Cobbold diversity measure is the most promising. However, two issues remain with this indicator: the use of a free sensitivity parameter that must be arbitrarily set depending on the studied system, and the absence of a standard for a classification and class similarities. This paper proposes a new reference-based diversity indicator to measure the interdisciplinarity of research papers that aims at solving these two issues, based on the expansion of the Gini coefficient to take class similarity into account. Since it contains no free parameter, our indicator can be used to compare different systems. In addition, our indicator provides a framework for the quantitative measurement of the quality of a classification.

Conference Topic
Indicators

Introduction
Since the publication of Gibbons et al. (1994) seminal book, interdisciplinary has been on the science policy agenda of several governments, research councils and universities. Indeed, many of these believe that, by combining the knowledge and methodologies of more than one discipline, greater scientific advances can be obtained (Frodeman, Klein, & Dos Santos Pacheco, 2017; Wagner et al., 2011). In this context, several bibliometric indicators have been developed to measure the extent to which a given body of research is interdisciplinary. However, the construction of appropriate indicators of interdisciplinarity—which can be considered as indicators of diversity—is a fair challenge, as exemplified by the several advanced indicators developed over the last decade (e.g. Stirling, 2007; Wang, Thijs & Glänzel, 2015; Yegros-Yegros, Rafols & D’Este, 2015; Mugabushaka, Kyriakou & Papazoglou, 2016; Zhang, Rousseau & Glänzel, 2016). Even the most advanced interdisciplinarity indicators—such as the Leinster-Cobbold and the Rao-Stirling—suffer from two main problems: the presence of free parameters, and the lack of standards for defining classes. This paper proposes a new interdisciplinary indicator which aims at correcting these two issues. It first describes the issue with existing indicators, and then presents of our new indicator, based on the expansion of the Gini coefficient, in which those issues are addressed.

Background
In its broadest sense, a diversity indicator is applied to an element, which is normally a set of different classes. Most diversity indicators were developed in disciplines such as biology and economics (Stirling, 2007). For example, an application of a diversity indicator would be if the element is a set of animals, and the classes are the animals’ species. Figure 1 shows an example
of an element, which could be represented as the set \( x = \{ \text{Class A} = 40 \%, \text{Class B} = 40 \%, \text{Class C} = 20 \% \} \), or \( x = \left[ \frac{2}{5}, \frac{2}{5}, \frac{1}{5} \right] \) for simplicity.

![Diagram of an element as a set of classes.]

Figure 1. Representation of an element as a set of classes.

In bibliometrics, the element is typically an article, or a set of articles, and classes are typically disciplines. The diversity of an element can then be calculated based on the bibliographic references, with each of the references belonging to one or multiple classes (or disciplines). Figure 2 shows, as an example, an article with five references: two in mathematics, two in chemistry, and one in physics.

![Diagram of an article as an element where the references’ domains are used as the studied set of classes.]

Figure 2. Representation of an article as an element where the references’ domains are used as the studied set of classes.

Once classes have been defined\(^1\), multiple properties can be taken into account for measuring the diversity of elements. As discussed by Stirling (2007), the three main properties are 1) the number of classes in the element (variety), 2) the evenness of the class distribution within the element (balance), and 3) the similarity between the classes (disparity). While considering these properties distinctively is crucial to understand the complex diversity of a single system, combining them in a single indicator is useful to compare different systems and correlate diversity to other variables on a larger scale.

While an abundance of diversity indicators that take into account only the balance (e.g., Simpson diversity, Shannon entropy and Gini coefficient) have been developed over the past decades, other indicators (e.g., Rao-Stirling diversity) have integrated disparity (Stirling, 2007). Among those new indicators, the Leinster-Cobbold measure of diversity (Leinster, & Cobbold, 2012) is one of the most promising, as it incorporates the three main properties: variety, balance and disparity. The Leinster-Cobbold formula is as follows:

\[^{1}\text{It is worth mentioning here that the notion of disciplines as classes can be problematic, since several disciplinary classifications cohabit in the literature, and ad-hoc definition of classes could in principle be just as valid as any other (Rafols, 2014).}\]
where $s_{ij}$ is the similarity between class $i$ and class $j$. It is through this similarity matrix that disparity is incorporated in the indicator. The strength of the Leinster-Cobbold indicator is that almost all relevant measures are derived from the sensitivity parameter $q$. However, because different values of $q$ provide significantly different results, and the ideal value of $q$ depends on the effect one wants to measure. For example, cases with a large number of rare classes require a higher value of $q$ to capture the effect of the transfer from one class to another. This creates a problem when we want to easily compare the diversity between systems from different regimes. This problem is made worse when we compare the diversity of a very large number of systems, like scientific articles. For this purpose, we need a new indicator that possesses the properties and advantages of the Leinster-Cobbold indicator without having to rely on a free parameter which optimal value depends on the studied system.

Any diversity indicator must satisfy two essential properties. First, the effective number property states that the diversity of a community of $N$ equally abundant, totally dissimilar classes must be $N$. The second essential property is Dalton’s transfer principle, which states that transferring unit abundance from a small class to a larger one should decrease diversity (Jost, 2009). The Leinster-Cobbold indicator does not respect Dalton’s transfer principle for every possible value of the similarity matrix $s_{ij}$. For some values of $s_{ij}$, adding a new class might decrease diversity. The new indicator we propose here adds a linear component to Dalton’s transfer principle, so that transferring a unit abundance from a small class to a larger one should decrease diversity linearly. Therefore, our indicator has no free parameter (e.g., the $q$ parameter in Leinster-Cobbold), which facilitates comparison between systems from different regimes. While the Gini coefficient (Gini 1997) already offers a linear behaviour when transferring units, it only measures the variety and balance of perfectly distinct classes, and therefore does not consider the similarity between classes (i.e., disparity), which makes the indicator weaker than Leinster-Cobbold. The proposed indicator is in fact an expansion of the Gini coefficient to include similarity.

The expansion of the Gini coefficient

The first step of the indicator is to renormalize the Gini coefficient so that it respects the effective number property. In other words, the Gini coefficient has a value between 0 and 1, while the renormalized coefficient should have a value between 1 and $N$. Let us consider a set of $N$ classes where $x_i$ is the share of elements of class $i$ in the set. For example, the values of $x_i$ for the case in Figure 1 would be $x_1 = 0.4, x_2 = 0.4$ and $x_3 = 0.2$. A renormalized Gini coefficient can be expressed as an ordered sum over all classes as shown in Equation (2).

$$
(2) \quad \sum_{i=1}^{N} (x_i - x_{i+1})^2 \quad \text{where } x_i > x_{i+1} \text{ and } x_i = 0 \text{ if } i > N
$$

One can see that Equation (2) satisfies both the effective number property and the linear version of Dalton’s transfer principle. Furthermore, its main advantage is that it can be interpreted as a linear combination of sets of equally abundant classes. In our example case, we express the Gini coefficient in terms of the diversity of a case where $x_1 = 0.5$ and $x_2 = 0.5$, and another
where \( x_1 = 0.33, \ x_2 = 0.33 \) and \( x_3 = 0.33 \). Indeed, assuming that the equally abundant set of the most frequent \( i \) classes has a diversity of \( D_i \), Equation (2) can be rewritten as:

\[
(3) \quad \sum_{i=1}^{N} i(x_i - x_{i+1})D_i
\]

In the case of totally dissimilar classes, \( D_i = i \) because of the effective number property, and we reobtain Equation (2). However, by changing this definition of \( D_i \) in Equation (3), we can obtain a version of the Gini coefficient that takes into account the similarity between classes. We can therefore attribute a value \( D_i \) for sets of equally abundant but not totally dissimilar classes, and our indicator can correct for disparity while considering the evenness of the distribution linearly. Here, the value of \( D_i \) will be defined by citations between classes. We consider the following hypothesis for \( D_i \): the proportion of all the references to a class \( A \) in articles of a class \( B \) are considered redundant information if class \( A \) was already considered in our set.

Let us consider a specific case. Because of the effective number property, we know that a set only comprised of class \( A \) must have a diversity of 1. If 10% of class \( B \) refers to class \( A \), then a set comprised equally of class \( A \) and class \( B \) should have a diversity of \( 1 + 1*(100\% - 10\%) = 1.9 \). However, for the process to be symmetrical, we also consider the references to class \( B \) in articles of class \( A \), and average the two values. Therefore, for \( i = 2 \) classes, the diversity \( D_i \) is equal to:

\[
(4) \quad D_{i=2} = 2 - \frac{1}{2} (r(A, B) + r(B, A))
\]

where \( r(A, B) \) is the proportion of the references of class \( A \) citing articles of class \( B \) and vice versa. If we consider more than two classes, we need to consider the average over all the possible combinations of classes, and the general formula becomes:

\[
(5) \quad D_i = N - \frac{1}{2} \sum_{j \neq k} r(j, k)
\]

where \( r(j, k) \) is the proportion of the references to class \( k \) in articles of class \( j \). Equation (6) shows that this can be expressed as the sum of all elements of a matrix if the diagonal elements are equal to 1 instead of the self-references of a class \( r(i, i) \). This apparent discontinuity is due to the effective number property that forces the diversity of a single class to be equal to 1.

\[
(6) \quad D_i = \sum_{j,k} M_{jk}
\]

\[
M_{jj} = 1
\]

\[
M_{jk} = -\frac{1}{2} r(j, k) \quad \text{if} \ j \neq k
\]

One should note that this formula does not respect Dalton’s transfer principle for all values of \( M_{jk} \). Indeed, a consequence of that principle is that adding a new class to an element should never decrease diversity. For example, if we consider the following \( M \):

\[
(7) \quad M = \begin{bmatrix}
1 & 0 & -\frac{1}{2} \\
0 & 1 & -\frac{1}{2} \\
-\frac{1}{4} & -\frac{1}{4} & 1
\end{bmatrix}
\]
We note a decrease in diversity from $D = 2$ for $x = [1 \, 1 \, 2]$ to $D = 1.5$ for $x = [1 \, 1 \, 3]$. This should not be too surprising since even the Leinster-Cobbold formula does not respect Dalton’s transfer principle for all values of the similarity matrix. Moreover, our indicator does not satisfy the Dalton’s transfer principle only in extreme cases where the sum of the relative references to a single class is greater than 1. The simple way to correct the formula is by modifying the diagonal of the matrix $M$ in the following way:

\[
M_{jj} = \max \left\{ 1, \sum_{k \neq j} r(k, j) \right\} = \omega_j
\]

This weight factor $\omega_j$ allows the indicator to respect Dalton’s transfer principle for all cases, and is justified by the fact that it corrects an inadequate classification. We consider that references from one class to another indicate their similarity and that those other classes have an intrinsic diversity value of 1 according to the effective number principle. Therefore, if the sum of references to a single class is greater than 1, then this class has technically been recognized by the other classes as too large and should theoretically be split in subclasses. Thus, even if this weight factor implies that the effective number principle is not satisfied, we can indirectly satisfy it by interpreting all the classes where $\omega_j$ is greater than 1 as a group of subclasses with a weight factor equal to 1.

We should note here that the absence of a recognized standard disciplinary classification creates a lot of possible choices defining classes (e.g., WoS subject categories, NSF journal classification, institutions, journals, etc.). Defining classes raises the question of the equilibrium between the number of classes and their similarity: is it better to have a large number similar classes, or a smaller number of dissimilar classes? We can solve this problem since we now have a quantitative bibliometric measure for the quality of a classification. A high number of classes provides a higher resolution of interdisciplinarity, which would increase the granularity of our indicator. Indeed, since all articles in the same class are considered to have the same characteristics, increasing the number of classes allows us to better account for differences between articles. However, classes should reflect reality regarding the effective number property, and we should therefore aim for all classes to have a value of $\omega_j = 1$. Since we already have multiple existing classifications to choose from, then a look at the average of $\omega_j$ can help us determine which one is the best. Therefore, to make our indicator as fine-grained as possible, we need a classification with the lowest average value of $\omega_j$ but the highest number of classes.

**Discussion and conclusion**

We developed a new indicator to measure the interdisciplinarity of an article or a group of articles based on the set of their references. It is an expansion of the Gini coefficient that takes into account the similarity between classes. Assuming the relative distribution of $n$ classes:

\[
\begin{bmatrix}
x_1 & x_2 & \ldots & x_n
\end{bmatrix}
\]

where $\sum_{i=1}^{N} x_i = 1$ and $x_i = 0$ if $i > N$

the diversity $D$ of our distribution is given by

\[
D = \sum_{i=1}^{N} i(x_i - x_{i+1}) \sum_{j=1, k=1}^{N} M_{jk}
\]

\[
M_{jk} = -\frac{1}{2} r(j, k) \quad \text{if } j \neq k
\]

\[
M_{jj} = \max \left\{ 1, \sum_{k=1, k \neq j} r(k, j) \right\} = \omega_j
\]
where $r(j,k)$ is the proportion of the references of class $j$ that cites articles of class $k$.

This formula offers a clear quantitative measure of the concept of diversity, defined as how a given element links—in this case, through references—classes that are not themselves strongly linked. Therefore, it is important that the measured element be independent from the process that establishes the similarity of the classes. Our diversity indicator is thus best suited for evaluating the diversity of a single article, or the articles of an author, an institution or a country. It possesses the inclusion of similarity effects of previous indicators such as the Leinster-Cobbold without relying on a free parameter, and provides a quantitative measure of the quality of a classification to establish a standard in that regard, while still satisfying all the essential properties of diversity indicators.

Further research will aim at comparing our new indicator with existing ones on empirical samples, reanalyzing patterns of interdisciplinarity as measured by our expanded Gini coefficient, and correlating its results with indicators of productivity, scholarly impact, or collaboration.

Acknowledgements

The authors wish to thank Diane Marie Plante, Philippe Mongeon, Jean Lagacé and Jonathan Laflamme Janssen for their numerous comments and suggestions.

References


Research preferences of the G20 countries: a bibliometrics and visualization analysis

Hu Zhigang¹ Lin Gege² Hou Haiyan³

¹huang@dlut.edu.cn
WISE Lab, Dalian University of Technology, Dalian (China)

²lingegenl@mail.dlut.edu.cn
WISE Lab, Dalian University of Technology, Dalian (China)

³htian@dlut.edu.cn
WISE Lab, Dalian University of Technology, Dalian (China)

Abstract
The purpose of this study is to reveal the differences both in research output and research preferences of the G20 countries. The research outputs of the nineteen G20 countries (excluding the European Union) are measured based on their publications indexed in Web of Science Core Collection. The research preferences of the G20 countries were investigated by comparing their research output in each research subject. Clustering method was then employed to classify the countries according to their research preferences. Nineteen countries are classified into four clusters. Countries assigned in same cluster are similar in distribution of research subjects. In the end, by VOSviewer, we showed research pattern of each cluster. For example, USA in Cluster A is characterized by the emphasis on the medical sciences; while China in Cluster C is characterized by paying more attention to physical sciences.

Conference Topic
Country-level studies

Introduction
The G20 is initially an international economic cooperation forum established on 1999. After almost twenty-year development, the G20 has become a major platform for international affairs and has played an increasingly important role in all kinds of global issues, including scientific research (Callaghan, Ghate, Pickford, & Rathinam, 2014). The G20 countries account for about 60% of the world's land area, 66.7% of the world population, and more than 90% of the sum of Gross Domestic Product. They are also the dominated producers of scientific research output, and served as the major engines to drive further development in science and technology.

However, the driving effects of the G20 countries vary sharply, both in their strength and their preferential research areas. Yang et al (2012) studied the research preference of the G7 countries, and found that life sciences are the main focuses of these developed country. Bouabid et al (2016) addressed the issue of the scientific collaboration between the G7 countries, theirs research showed that the G7 countries had intensive intra-collaboration activities. Thomson Reuters (now named Clarivate Analytics) investigated the research and innovation performance of the G20 in a report in 2014 (Thomson Reuters, 2014), and list the amount and world share of each G20 country’s publications in the select OECD research fields. Hu et al (2017) explored research preferences of the provinces of China by the method of cosine similarity and hierarchical clustering, and mapped different province’s research hotspots using VOSviewer (Hu, Guo, & Hou, 2017). Almeida et al (2009) performed an analysis on the way how the European countries are clustered according to their similarity (Almeida, Pais, & Formosinho, 2009).
In this research, we systematically examined research output and preference of the G20 countries. The following questions will be addressed below: (a) how great is the differences among the G20 countries in their research output? (b) what is the research preference for each country? (c) which countries have the similar/different research preferences? (d) does a country’s economic level have influence on its research preferences?

Data and methods

Data collection

The G20 countries’ publications were retrieved in Web of Science™, one of the world’s most comprehensive bibliometrics database. WoS Core Collection is consisted of three core journal databases, namely Science Citation Index (SCI), Social Sciences Citation Index (SSCI), and Arts & Humanities Citation Index (A&HCI). Together they cover more than 10,000 journals of the research subjects of natural sciences, social sciences and humanities. WoS Core Collection also integrated two important conference databases, CPCI-S and CPCI-SSH. All these journals and proceeding books are assigned into one or more research subjects, such as Engineering Electrical Electronic, Materials Science, Oncology, etc. In WoS, we retrieved the amount of publications of the G20 countries for each research subject in 2015. To exclude non-research articles, only Article, Proceedings Paper and Review are included, and all the other data type, such as Letter, Note, and Edictal Martial are ignored. It also should be noted that the WoS subject categories overlap in coverage and therefore some publications have been double counted.

Similarity Measurement between G20 countries

Two countries are similar in research preference if their cosine distance is close(Singhal, 2001). Since this study is intended to compare the research preference instead of absolute research output, the cosine distance is more appropriate than Euclidean distance. After calculating the similarity of the G20 countries, we further classified them into four clusters using the method of hierarchical clustering. Hierarchical clustering shows not only the result of clustering but also the clustering process through a tree diagram. The most similar two countries are merged firstly. The process continues until all nineteen countries are merged(B & Suzuki, 2015).

Research Hotspot of the G20 countries

To illustrate research preference vividly, we draw research hotspots for several typical countries using the visualization technology in the end. VOSviewer, a software tool developed by Van Eck and Waltman for constructing bibliometrics networks (Eck & Waltman, 2013, 2014; Hu et al., 2017; van Eck & Waltman, 2010), is employed in this research. Figure 1 is the global map of research subjects, which provides a foundational research hotspot in the world today. This map is derived from the official website of VOSviewer. In this map, all the Web of Science’ research subjects are located based on their citation relations to each other. Five research fields are clustered and assigned with distinguished colours. They are the field of Social Sciences and Humanities in red, the field of Health Sciences in green, the field of Life Sciences in yellow, the field of Physical Sciences and Engineering in blue, and the field of Mathematics and Computational Sciences in purple. Of particular note is that Health Sciences and Life Sciences are allied research fields overlapping with each other. Health Sciences mainly refer to the subjects related with medicine, nursing and pharmaceutical science; while Life Sciences mainly refer to agricultural and environmental sciences.
Based on the research map, we generated network maps for different G20 countries by keeping the layout of map constant. The size of each node was redrew according to the amount of publications of and the country in the corresponding research subject. The research network is able to show the relatedness of nodes and the weight of the item in the network. The VOSviewer maker notes that for some items the label may not be visible. This is done in order to avoid overlapping labels (Eck & Waltman, 2013).

Cluster analysis of the G20 countries

Research output of the G20 countries

The publication amount of the G20 countries in 2015 is listed in Table 1. The United States, not surprisingly, ranks first with about 0.45 million publications. China ranks second with about 0.36 million publications. They are the undisputed leaders in research output, far beyond anyone else in the G20 countries. The sum of publications from the third ranked UK, the fourth ranked Germany and the fifth ranked Japan are still less than China’s research output, let alone the US. Obviously, the G20 countries vary greatly in terms of counts of publications. The research output of the US is almost 81 times that of Indonesia, the least producer in the G20.

Generally, a country's scientific research output is proportional to its economic level. The USA and China, which produced the most research publication, are also the bigger economic powers. UK, German and Japan contribute the third to fifth research publications, their GDP are also ranked from the third to fifth, just in reversed order. France, India and Canada ranks the sixth to eighth both in research publications and GDP. The top 13 research producers are exactly the same with the richest countries, only in somehow different sequence.

Table 1 Publication counts of G20 countries in Web of Science

<table>
<thead>
<tr>
<th>Country</th>
<th>Publications</th>
<th>GDP Rank</th>
<th>Country</th>
<th>Publications</th>
<th>GDP Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>447,685</td>
<td>1</td>
<td>11</td>
<td>64,178</td>
<td>11</td>
</tr>
<tr>
<td>China</td>
<td>357,727</td>
<td>2</td>
<td>12</td>
<td>48,782</td>
<td>9</td>
</tr>
<tr>
<td>UK</td>
<td>125,333</td>
<td>5</td>
<td>13</td>
<td>45,013</td>
<td>12</td>
</tr>
<tr>
<td>Germany</td>
<td>125,139</td>
<td>4</td>
<td>14</td>
<td>32,309</td>
<td>18</td>
</tr>
</tbody>
</table>
The scatter plot chart in Figure 2 shows the correlation between the research output and the GDP of the G20 countries. For each country, coordinate X represents its GDP in 2015, and Y represents its publication count. It shows that a country’s GDP is significantly \((R^2=0.9517)\) correlated with its research output.

![Fig. 2 The relationship between GDP and the amount of publications (2015)](image)

Publications of research subjects

In Table 2, the publication amounts of 252 research subject are listed. “Engineering Electrical Electronic” is the largest subjects. It involved 163,833 publications, and was 68.7% higher than the second-placed “Materials Science Multidisciplinary”. The subjects of “Physics Applied”, “Chemistry Multidisciplinary” and “Chemistry Physical” rank third to fifth respectively. The top 5 research subjects contributed more than 1/4 of all publications. Correspondingly, the publications in the subjects of “Literature African Australian Canadian” and “Poetry”, are only less than 200, merely account for 0.0008%.

<table>
<thead>
<tr>
<th>Web of Science Categories</th>
<th>Pub.</th>
<th>Web of Science Categories</th>
<th>Pub.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering Electrical Electronic</td>
<td>163,833</td>
<td>Literature Slavic</td>
<td>580</td>
</tr>
<tr>
<td>Materials Science Multidisciplinary</td>
<td>97,143</td>
<td>Psychology Psychoanalysis</td>
<td>566</td>
</tr>
<tr>
<td>Physics Applied</td>
<td>77,730</td>
<td>Andrology</td>
<td>502</td>
</tr>
<tr>
<td>Chemistry Multidisciplinary</td>
<td>67,224</td>
<td>Literature German Dutch Scandinavian</td>
<td>463</td>
</tr>
<tr>
<td>Chemistry Physical</td>
<td>59,842</td>
<td>Literature British Isles</td>
<td>421</td>
</tr>
<tr>
<td>Computer Science Theory Methods</td>
<td>59,024</td>
<td>Dance</td>
<td>367</td>
</tr>
<tr>
<td>Multidisciplinary Sciences</td>
<td>56,745</td>
<td>Literature American</td>
<td>325</td>
</tr>
<tr>
<td>Biochemistry Molecular Biology</td>
<td>53,522</td>
<td>Folklore</td>
<td>306</td>
</tr>
<tr>
<td>Optics</td>
<td>48,839</td>
<td>Literature African Australian Canadian</td>
<td>154</td>
</tr>
</tbody>
</table>

Table 2 Publications of Research subjects in Web of Science (2015)
Research preferences of the G20 countries

By listing top research subjects of each G20 country, we uncovered the research preferences in Table 3. In 2015, the USA researchers published the most publications in the subjects of “Engineering Electrical Electronic”, “Multidisciplinary Sciences”, “Materials Science Multidisciplinary”, “Biochemistry Molecular Biology” and “Neurosciences”. Chinese researchers did well in the subjects of “Engineering Electrical Electronic”, “Materials Science Multidisciplinary”, “Chemistry Multidisciplinary”, “Physics Applied” and “Chemistry Physical”. Compared with the USA and China, UK paid more attention to “Astronomy Astrophysics”, and Indonesian preferred published more papers in the subject of “Computer Science Information Systems”, “Computer Science Theory Methods” and “Telecommunications”.

Table 3 Preferential research subjects of each G20 countries in Web of Science TM (2015)

<table>
<thead>
<tr>
<th>USA</th>
<th>China</th>
<th>UK</th>
<th>...</th>
<th>Indonesia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering Electrical Electronic (28547)</td>
<td>Engineering Electrical Electronic (41016)</td>
<td>Engineering Electrical Electronic (6782)</td>
<td>...</td>
<td>Engineering Electrical Electronic (984)</td>
</tr>
<tr>
<td>Multidisciplinary Sciences (18388)</td>
<td>Materials Science Multidisciplinary (3912)</td>
<td>Multidisciplinary Sciences (5567)</td>
<td>...</td>
<td>Computer Science Information Systems (575)</td>
</tr>
<tr>
<td>Materials Science Multidisciplinary (16104)</td>
<td>Chemistry Multidisciplinary (21104)</td>
<td>Materials Science Multidisciplinary (3866)</td>
<td>...</td>
<td>Physics Applied (549)</td>
</tr>
<tr>
<td>Biochemistry Molecular Biology (15395)</td>
<td>Physics Applied (19927)</td>
<td>Astronomy Astrophysics (3706)</td>
<td>...</td>
<td>Computer Science Theory Methods (390)</td>
</tr>
<tr>
<td>Neurosciences (14131)</td>
<td>Chemistry Physical (17192)</td>
<td>Neurosciences (3648)</td>
<td>...</td>
<td>Environmental Sciences (325)</td>
</tr>
<tr>
<td>Physics Applied (14038)</td>
<td>Energy Fuels (14283)</td>
<td>Biochemistry Molecular Biology (3440)</td>
<td>...</td>
<td>Telecommunications (248)</td>
</tr>
<tr>
<td>Oncology (13375)</td>
<td>Optics (13162)</td>
<td>Physics Applied (3227)</td>
<td>...</td>
<td>Computer Science Interdisciplinary Applications (240)</td>
</tr>
<tr>
<td>Public Environmental Occupational Health (12049)</td>
<td>Computer Science Theory Methods (12787)</td>
<td>Computer Science Theory Methods (2976)</td>
<td>...</td>
<td>Engineering Industrial (234)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Dance (29)</td>
<td>Literature German Dutch Scandinavian (0)</td>
<td>Literature American (6)</td>
<td>...</td>
<td>Sport Sciences (0)</td>
</tr>
<tr>
<td>Literature Slavic (20)</td>
<td>Literature Romance (0)</td>
<td>Literature Slavic (6)</td>
<td>...</td>
<td>Transplantation (0)</td>
</tr>
</tbody>
</table>

The similarity of the G20 countries in research preferences

The cosine distances between each two countries were calculated based on the data in Table 3. As it shown in Table 4, the resulting adjacency matrix represents the level of similarity between every two countries. For example, the cosine similarity between Australia and China is 0.7991; while that between Australia and France is 0.8906. It means Australia is more similar with France than China in the distribution of research subjects.

Table 4 Adjacency matrix of G20 countries’ cosine similarity in research areas

<table>
<thead>
<tr>
<th>Country</th>
<th>Australia</th>
<th>China</th>
<th>France</th>
<th>India</th>
<th>...</th>
<th>Japan</th>
<th>Russia</th>
<th>UK</th>
<th>USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>0.7991</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>France</td>
<td>0.8906</td>
<td>0.9019</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>India</td>
<td>0.7546</td>
<td>0.9224</td>
<td>0.8864</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5 lists the country pairs with the most and the least similarity. France and Germany, two adjacent countries located in the Western Europe, are closest in the vector space of research subjects. Their similarity is 0.9797. This similarity between them also show in Almeida's research (Almeida et al., 2009). The next closest countries pairs are between UK and USA, both important developed countries in the word with similar economic status and structure. The third pairs are Canada and USA. Although the absolute count of publications varies a lot between Canada and USA, both of them have a homogeneous disciplinary structure and research preferences. The most dissimilar countries pairs are Argentina and Indonesia. Their similarity is only 0.5005.

Table 5 The cosine distance between G20 countries (in descending order)

<table>
<thead>
<tr>
<th>Country A</th>
<th>Country B</th>
<th>Cosine distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>Germany</td>
<td>0.9797</td>
</tr>
<tr>
<td>UK</td>
<td>USA</td>
<td>0.9760</td>
</tr>
<tr>
<td>Canada</td>
<td>USA</td>
<td>0.9736</td>
</tr>
<tr>
<td>Canada</td>
<td>UK</td>
<td>0.9683</td>
</tr>
<tr>
<td>Australia</td>
<td>Canada</td>
<td>0.9653</td>
</tr>
<tr>
<td>Australia</td>
<td>UK</td>
<td>0.9634</td>
</tr>
<tr>
<td>France</td>
<td>Italy</td>
<td>0.9603</td>
</tr>
<tr>
<td>Germany</td>
<td>Italy</td>
<td>0.9562</td>
</tr>
<tr>
<td>Germany</td>
<td>Japan</td>
<td>0.9561</td>
</tr>
<tr>
<td>Germany</td>
<td>USA</td>
<td>0.9547</td>
</tr>
<tr>
<td>....</td>
<td>....</td>
<td></td>
</tr>
</tbody>
</table>

Cluster analysis of the G20 countries

The dendrogram of the cluster analysis is shown in Fig. 3. This is a pictorial representation of the data structure, indicating the merging objects and the merging distances.

Cluster A is composed of eight countries: France, Germany, Italy, USA, UK, Canada, Australia and Turkey. Besides Turkey, all the others are all developed countries and located in Europe and North America.

Cluster B is composed of four countries: Brazil, Mexico, Argentina and South Africa. They are all former European colonies, located in either Latin America or Africa.
Cluster C consists of six countries: Japan, South Korea, China, Saudi Arabia, India, and Russia. These six countries are mainly Asian countries, with different cultures and languages from the West.

Cluster D contains only one country, namely, Indonesia, the smallest research producer in the G20. Indonesia is famous for its superior natural conditions and abundant marine resources. It contains more than seventeen thousand islands. The electronics industry is Indonesia’s most important engine of economic development.

At a highest level, the eleven countries in Clusters A and B merged into Group I; while the other seven countries in Cluster C and D merged into Group II. Group I is composed of developed countries mainly from the West. Compared with those in Group I, the countries in Group II are almost located in Asia, and most of them are developing countries.

Although economic level has a significance effect on discipline structure and subject development, geographical locations might have a greater influence on research preferences. For example, Japan and South Korea, two developed countries, both belong to Group II instead of Group I.

Research mapping of the clusters

Each cluster is featured by its unique research preference. In the following section, we selected one typical country of each cluster. They are the USA in Cluster A, Brazil in Cluster B, China in Cluster C, and Indonesia in Cluster D. We will map their research hotspots using VOSviewer.

a. Research map of the USA

The research map of the USA is shown in Figure 4. Health Sciences is the hottest field in the USA. Specifically, the researchers of the USA focus on the subjects of “Biochemistry Molecular Biology”, “Neurosciences”, “Oncology”, “Public Environmental Occupational Health” and “Surgery”, “Cell Biology”, etc.

Another field preferred by the USA are Physical Sciences and Engineering. In this field, “Materials Science Multidisciplinary”, “Physics Applied”, “Chemistry Physical”, “Chemistry Multidisciplinary” and “Optics” are the most dominant subjects. Besides, the USA also an
important producer of the research papers in the subject of “Engineering Electrical Electronic” and other subjects in the field of Computational Sciences.

Compared with the countries in other clusters, the USA also contribute a lot in the field of Social Sciences and Humanities, in which field the most primary subjects are “Psychiatry”, “Economics”, “Education Educational Research”, “Psychology Clinical”, etc. The USA is less productive in the field of Life Sciences, which is mainly involved with Agriculture sciences and Environmental Sciences.

The USA shows a typical research maps for North American and European countries, which was called “Western Model” by a report published in 1997 (REIST-2, 1997). In this report, “Western Model” is defined by focusing on the research in clinical medicine and biomedical sciences.

b. Research map of Brazil

Brazil is featured by its advantage in the field of Life Sciences. Brazil mainly focus on the subjects of “Plant Sciences”, “Agronomy”, “Food Science Technology”, “Environmental Sciences” and “Agriculture Multidisciplinary”. Abundant natural resources and agricultural production provide an essential necessity and condition for Brazil and other countries in Cluster B.

Health Sciences is also a preferential research field of Brazil. In this field, the subjects of “Biochemistry Molecular Biology”, “Public Environmental Occupational Health”, “Pharmacology Pharmacy”, “Dentistry Oral Surgery Medicine”, “Neurosciences” are the most primary subjects.

In the research fields of Physical Sciences and Engineering or Mathematics and Computational Sciences, Brazil also had good performance. But The Brazilian researchers seemingly pay less attention to Social Sciences and Humanities, in which field an only primary subject of the Brazil is Psychiatry.
The Brazil’s research map is in accordance with the result of the report (REIST-2, 1997). As we find, Brazil is still following the “Bio-Environmental Model” with Biology, Earth and Space sciences in the main focus. The finding is also confirmed by a study of GLÄNZEL (GLÄNZEL, LETA, & THIJS, 2006).

c. Research map of China

China is the typical developing country. Chinese researchers have been more active in the field of Physical Sciences and Engineering, as well as Mathematics and Computational Sciences. The emphasizing on physical sciences is also stated by previous research (Fu, Chuang, Wang, & Ho, 2011). In comparison, Health sciences and Life sciences are not so preferential in China as the USA and Brazil.

“Materials Science Multidisciplinary”, “Chemistry Multidisciplinary” and “Physics Applied”, in the field of Physical Sciences, “Engineering Electrical Electronic”, “Computer Science Theory Methods” and “Automation Control Systems” in the field of Mathematics and Computational Sciences are the most productive subjects in China.

In the field of Health Sciences, China focuses on the subjects of “Oncology”, “Biochemistry Molecular Biology” and “Medicine Research Experimental”. In the field of Life Sciences, “Plant Sciences”, “Agronomy”, “Food Science Technology” are the most preferential subjects. China’ mainly concerned with the subjects of “Social Sciences Interdisciplinary”, “Education Educational Research” and “Folklore” in the field of Social Sciences and Humanities.

The heavy emphasis on classical sciences (chemistry, physics, engineering etc) in China was similar three decades ago. Although China has increased investment in and policy support for emerging industries of strategic importance, such as biotechnology and pharmaceutics, their development are currently facing key issues such as lack of innovation of the driving force and endogenous growth fatigue.
d. Research map of Indonesia


The field of Physical Sciences and Engineering is another important research fields for Indonesia, “Physics Applied”, “Materials Science Multidisciplinary” and “Energy Fuels” are the most dominant subjects.

Indonesia also focused on the field of Social Sciences Interdisciplinary, such as “Social Sciences Interdisciplinary”, “Business”, “Management”, “Education Educational Research”, etc. Compared with Health Sciences, Indonesia prefers to conduct research in the field of Life Sciences.

In the end, we drew radar charts for each country above, and reveal the difference in their research maps, as shown in Figure 8. By comparing the shape of radar charts, we are able to
distinguish the research spotlights of each country or cluster more clearly. The different clusters’ different characteristics in research preferences could be identified obviously.

Conclusions
We compared the research performances of the nineteen G20 countries in this research. As we found, different countries vary greatly not only in research outputs, but also in research preferences. The research publications of the USA could be 80 times as that of Indonesia. The difference in terms of research preferences is as well. By clustering method, the nineteen G20 countries are classified into four clusters firstly and then two groups at a higher level. Each cluster is featured with a particular research preference. For example, the countries in Cluster A, e.g., the USA, prefer the research area of Health Sciences; while those in Cluster C, e.g., China, put a greater emphasis on the research area of Physical Sciences and Engineering. The research also found that there is significant correlation between a national research performance and its economic level or geographic location. If some countries have similar economic levels or their locations are close to each other, their research preferences tend to be similar. The developed countries conduct more research in the biology and medical sciences; while the developing countries’ emphasises go to physical sciences and computer sciences. Using VOSviewer and Radar Charts, we provide research maps of four typical countries in each cluster. It allows us compare research preferences of different clusters more intuitively.

Acknowledgments
The research was supported by the Natural Science Foundation of China (NSFC) under Grant 71503031 and China Postdoctoral Science Foundation under Grant 2016M591435.

Fig. 8 Radar charts of the selected four countries in different clusters
References


Bibliometric Study of Interdisciplinary Relations of Converging Technologies (Nano-Bio-Info-Cogno)

Hamid R. Jamali1  Saeid Asadi2  Ghasem Azadi-Ahmadabadi3

1 h.jamali@gmail.com
Charles Sturt University (Australia)
Corresponding author

2 s.asadi@shahed.ac.ir
Shahed University (Iran)

3 azadi_gh@yahoo.com
Kharazmi University (Iran)

Abstract
This study investigates the interdisciplinary relations of nanotechnology (Nano), biotechnology (Bio), information technology (Info), and cognitive science (Cogno) (together known as NBIC converging technologies) at the science level using a range of different bibliometric techniques and measures. The study applied journal citation, author citation, keyword analysis, and authorship analysis on all Iranian NBIC international articles (2001-2015). Spanning tree, inclusion index, import and export of concepts, and inward and outward interactions were used to measure the interdisciplinary relations. The results showed that Bio and Nano had the highest level of interdisciplinary relations in terms of journal citations, shared keywords, and the exchange of authorship. On the other hand, Cogno had no journal citation relation with Bio, and it had very low exchange of concepts with Bio. Cogno also had the weakest link to Nano in terms of authorship as well as author citations. The study suggests using a combination of techniques and measures for the study of interdisciplinarity and convergence.

Conference Topic
Science communication; Participation in science

Introduction
Interdisciplinary relations are interactions between two or more disciplines. These interactions could be as simple as the exchange of ideas or as complex as mutual integration of methods, terminologies, data and epistemology of disciplines. Interdisciplinary relations play an important role in the progress of science and the study of interdisciplinary relations has gained a great deal of importance in recent years. In the bibliometric tradition, interdisciplinarity has remained a difficult issue despite its high policy-relevance (Wagner et al., 2011). However, access to high-volume and high-quality data sets of scientific output (e.g., publications, patents, grants) and computers and algorithms capable of handling this enormous stream of data has made it possible for bibliometricians to deeply study the phenomenon of interdisciplinarity (Borner, Maru and Goldstone, 2004).

Interdisciplinary research is considered a kind of convergence phenomenon (Jeong, Kim and Choi, 2015). There are four areas that their convergence is particularly considered important, especially in terms of technology convergence. They include nanotechnology, biotechnology, information technology, and cognitive science (together known with the acronym of NBIC) (Jeong, Kim and Choi, 2015). This technological convergence might manifest itself in the early stages as the interdisciplinary relations and interactions among these four fields. Technology convergence has different levels that starts with convergence in science and moves to convergence in technology, market and then in industry (Jeong, Kim and Choi, 2015). Although, individual fields from NBIC, especially nanotechnology and nanoscience, have received considerable attention in bibliometric studies, there is not much research on the
interdisciplinary relations of these four areas. It is critical to know how these converging fields rely on each other, how similar they are and how much scientific exchange and interactions occur among them. All of these four fields have been especially emphasised in the Iranian Scientific Roadmap and the Iranian government has invested in research in these areas. This article aims to investigate the convergence of these four fields at the science level by focusing on the relations between them. The focus of the article is on Iranian international publications in these areas. There are different bibliometric techniques for studying interdisciplinary relations and this article uses a combination of techniques including journal citations, author citations and keyword analysis.

Literature review

Techniques and methods for measuring similarity and interdisciplinary relations in science have been laboriously developed over the past several decades using bibliometric analysis (Rafols, Porter and Leydesdorff, 2010). There have been already a few reviews and discussions of the methods used for this purpose including Van Raan (1999) and Wagner et al. (2011). Generally, we can divide studies on the interdisciplinary relations of scientific fields into five groups based on the techniques they have used.

The first group include studies that have used citation analysis. These studies focus on the cross references between scientific fields with the assumption that if two fields of science cite publications of one another, they are related. For instance Goldstone and Leydesdorff (2006) studied how much cognitive science cite other fields of science and how much it is cited by other fields. The results showed that cognitive science cited neurology (13.4%) more than other fields and it received the highest rate of interdisciplinary citations from computer and artificial intelligence (15.2%). In another study Porter and Youtie (2009) studied interdisciplinarity of nanotechnology and found that nanotechnology had some level of convergence with medicine, biotechnology, neurosciences, physics and maths. In a more recent study on nanotechnology and nanoscience, Stopar et al. (2016) found out that nanosciences showed characteristics of a distinct discipline.

The second group are studies that focus on co-classification analysis which looks at the co-occurrence of different subject-classification headings assigned to research publications. An example from this group is the study by Rafols and Meyer (2010) on bionanoscience.

The third group have used co-word analysis. Milojević (2012) in his study of nanotechnology found that some 85% of Nano research was multidisciplinary. Wang, Notten and Surpatean (2013) in their study of nanotechnology also found that nanotechnology has developed to a relatively mature stage and has become a standardized and codified technology.

The fourth group have used co-authorship analysis techniques. Bergmann et al. (2016) investigated how interdisciplinary collaboration in cognitive science was and they found a high level of interdisciplinary collaboration in cognitive sciences.

The fifth group are those that have applied co-citation and bibliographic coupling analysis techniques. For instance, Bassecoulard, Lelu and Zitt (2007) used bibliographic coupling analysis for nanosciences and found out that observed moderate multi-disciplinarity is based on a real inter-disciplinarity at the paper level. Chen et al. (2015) used a co-citation matrix to study the interdisciplinary evolution of biochemistry and molecular biology. Their study showed that interdisciplinarity evolves mainly from neighbouring fields to distant cognitive areas and provides evidence of an increasing tendency of biochemistry and molecular biology researchers to cite literature from other disciplines.

Overall, there is a wide range of research on interdisciplinary relations as well as on identifying converging fields (such as Buter, Noyons and Van Raan, 2010). Several studies in the past have focused on nanotechnology and nanoscience and a few on cognitive sciences. However, there
seems to be no study on the relations between all of the four areas of NBIC and the present study contributes in this area.

**Methods and measures**

Data for the Iranian articles were obtained from Scopus database. Any English article with at least one author affiliated to an Iranian institute was retrieved in four fields of nanotechnology (Nano), biotechnology (Bio), information technology (Info), and cognitive science (Cogno) for the period 2001-2015. The data were searched and obtained in 6 September 2016 as batches of 2000 records (the maximum number Scopus allows downloading). The year of publication was used to break the search results into smaller sets in order to be able to download the records. They were then imported into a database for the analysis. Table 1 shows the number of Iranian records for each subject area.

<table>
<thead>
<tr>
<th>Field</th>
<th>No of publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nanotechnology</td>
<td>21,393</td>
</tr>
<tr>
<td>Biotechnology</td>
<td>27,578</td>
</tr>
<tr>
<td>Information technology</td>
<td>16,358</td>
</tr>
<tr>
<td>Cognitive science</td>
<td>3,535</td>
</tr>
</tbody>
</table>

Four analysis techniques were used: journal citations, author citations, keyword analysis and authorship.

- **Journal citations**: citations to journals by publications of each field were analyzed and then the amount of overlap between the cited journals (journals jointly cited by each pair of fields) was calculated.
- **Author citations**: citations to authors by publications of each field were analyzed and then the amount of overlap between the cited authors (authors jointly cited by each pair of fields) was calculated.
- **Keywords analysis**: VOSviewer (vosviewer.com) was used to do a keyword analysis on the titles of the articles for each field. For each field, keywords with at least five occurrences were extracted and then the amount of keywords shared by each pair of fields was calculated. VOSviewer was used to extract the keywords from titles of each field. The results were presented to a panel of experts who identified common terms to be removed. The remaining sets were used for comparison of the four fields as pairs or triads.
- **Authorship analysis**: the number of authors who have published articles in each pair of fields were calculated. Author analysis was only done on Iranian authors and foreign co-authors were removed from the sets. Affiliations were used to do some control for homonyms.

Four measures were used in this study: inclusion index, import/export of concepts, inward/outward interactions and spanning tree.

- **Spanning tree**: spanning tree calculation was used to show the strongest (maximum spanning tree) and weakest (minimum spanning tree) ties between the subject fields.
- **Inclusion index**: inclusion index is a simple measure for asymmetric similarity. It was used because among three indexes that have been suggested for measuring similarity of
documents (including Salton’s Cosine formula, the Jaccard Index, and the Inclusion Index), inclusion index reveals better results (Sternitzke & Bergmann, 2009). In asymmetric similarity, the similarity of set A to set B is different from the similarity of set B to set A (Qin, 2000). Inclusion index here indicate the extent by which keywords from one field have been used in a second field. The formula used here was $100 \times \frac{A}{M}$ in which A is the number of keywords common in both fields of B and C, and M is the number of keywords from field B that do not exist in field C.

- Import and export of concepts: this measures (used only for keyword analysis) shows how many concepts from one field are used in other fields (exporting) and how many concepts from other fields are used in the given field (importing). Concepts here are represented with keywords.
- Inward and outward interactions: this measure (used only for author analysis) was used to see the percentage of authors of a given field that have authored articles in other fields (outward interaction) and the percentage of authors from other fields that have authored articles in the given field (inward interactions).

**Findings**

**Journal citations**

Table 2 shows the amount of overlap in each pair of fields in their citations to journals. For instance the first row shows the overlap between Nano and Cogno. It shows that 17 journals have been cited by the articles of both fields. There have been 223 citations to these 17 journals in Nano articles, accounting for 9.9 per cent of all of citations. Cogno articles included 4,424 citations (12%) to these 17 journals. The largest number of journals jointly cited by two fields belongs to Nano and Bio with 297 journals cited by both fields. While these journals accounted for 65.4 percentage of citations from Bio articles, they only were 18 per cent of citations from Nano articles. The smallest overlap is between Bio and Cogno (no overlap) and then between Info and Bio, and Info and Cogno respectively.

Figure 1 shows the weighted graph of the spanning tree representing the journal citation similarities between the four fields. The numbers shown for edges are percentages. The number inside nodes are the total number of journals cited in the articles of that field. The maximum spanning tree is shown with dashed line and the minimum spanning tree is shown with dotted line. The strongest relation is between Bio and Nano where 65.4 per cent of citations in Bio articles are to sources that have been cited in Nano articles too. The weakest relation is between Bio and Cogno where Bio has no common cited sources with Cogno articles.

<table>
<thead>
<tr>
<th>Combination</th>
<th>No of journals cited by both fields</th>
<th>First field</th>
<th>No citations of journals</th>
<th>% of citations</th>
<th>Second field</th>
<th>No citations of journals</th>
<th>% of citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nano &amp; Cogno</td>
<td>17</td>
<td>Nano</td>
<td>223</td>
<td>9.9</td>
<td>Cogno</td>
<td>4,424</td>
<td>12</td>
</tr>
<tr>
<td>Nano &amp; Bio</td>
<td>297</td>
<td>Nano</td>
<td>40,521</td>
<td>18</td>
<td>Bio</td>
<td>168,265</td>
<td>65.4</td>
</tr>
<tr>
<td>Nano &amp; Info</td>
<td>101</td>
<td>Nano</td>
<td>55,102</td>
<td>24.6</td>
<td>Info</td>
<td>6,081</td>
<td>4.4</td>
</tr>
<tr>
<td>Info &amp; Cogno</td>
<td>13</td>
<td>Info</td>
<td>1,612</td>
<td>1.2</td>
<td>Cogno</td>
<td>1,588</td>
<td>4.4</td>
</tr>
<tr>
<td>Info &amp; Bio</td>
<td>35</td>
<td>Info</td>
<td>6,397</td>
<td>4.7</td>
<td>Bio</td>
<td>3,787</td>
<td>1.5</td>
</tr>
<tr>
<td>Cogno &amp; Bio</td>
<td>6</td>
<td>Cogno</td>
<td>7</td>
<td>0.01</td>
<td>Bio</td>
<td>7</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Inclusion index was calculated for overlap in citations to journals. Table 3 shows that for instance the similarity of journal citations in Nano articles to Bio articles (2nd row) is 18.6%, while the similarity of journal citations in Bio articles to Nano articles is higher (22.1%). The smallest number is the similarity of journal citations in Cogn articles to journal citations in Bio articles.

Table 3. Inclusion index for overlap of citations to journals

<table>
<thead>
<tr>
<th>Combinations</th>
<th>First field</th>
<th>Inclusion index (%)</th>
<th>Second field</th>
<th>Inclusion index (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nano &amp; Cogno</td>
<td>Nano</td>
<td>1.06</td>
<td>Cogno</td>
<td>5.3</td>
</tr>
<tr>
<td>Nano &amp; Bio</td>
<td>Nano</td>
<td>18.6</td>
<td>Bio</td>
<td>22.1</td>
</tr>
<tr>
<td>Nano &amp; Info</td>
<td>Nano</td>
<td>6.3</td>
<td>Info</td>
<td>10.8</td>
</tr>
<tr>
<td>Info &amp; Cogno</td>
<td>Info</td>
<td>1.4</td>
<td>Cogno</td>
<td>4</td>
</tr>
<tr>
<td>Info &amp; Bio</td>
<td>Info</td>
<td>3.7</td>
<td>Bio</td>
<td>2.6</td>
</tr>
<tr>
<td>Cogno &amp; Bio</td>
<td>Cogno</td>
<td>0.45</td>
<td>Bio</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Author citations
Another way of measuring the relation between two fields or their similarity is to look at their overlap in citations to authors. Table 4 shows the overlap between each pair of fields in their author citations. For instance there were 1,494 authors that were cited by articles in both Nano and Cogn articles. In Nano articles, 32,391 citations (3.4% of all of citations in Nano articles) were made to these 1,494 authors, while this figure for Cogn science articles was 45,334 citations (23.5% of all of citations in Cogn articles). The strongest relation is where 72.5 percent of citations in Nano articles were made to 9,479 authors who also were cited by articles in Bio. The data are also illustrated in the spanning tree graph (Figure 2). The number inside nodes are the total number of citations to authors in the articles of that field.

Table 4. Overlap of citations to authors in each pair of subjects

<table>
<thead>
<tr>
<th>Combinations</th>
<th>No of authors</th>
<th>First field</th>
<th>No of citations</th>
<th>Percentage of citations to First field</th>
<th>Second field</th>
<th>No of citations</th>
<th>Percentage of citations to Second field</th>
</tr>
</thead>
</table>

Figure 1. Spanning tree of the relations of four fields based on journal citations
Inclusion index for author citations is presented in Table 5. We can see that the similarity of Nano articles to Cogno articles in terms of citing the same authors is 4.6 per cent while Cogno articles are only 3.3 per cent similar to Nano in this regard. The highest similarity is the similarity of Info articles to Bio articles (9.4%).

Table 5. Inclusion index for overlap of citations to authors

<table>
<thead>
<tr>
<th>Combinations</th>
<th>First field</th>
<th>Inclusion index (%)</th>
<th>Second field</th>
<th>Inclusion index (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nano &amp; Cogno</td>
<td>Nano</td>
<td>4.6</td>
<td>Cogno</td>
<td>3.3</td>
</tr>
<tr>
<td>Nano &amp; Bio</td>
<td>Nano</td>
<td>1.6</td>
<td>Bio</td>
<td>1.6</td>
</tr>
<tr>
<td>Nano &amp; Info</td>
<td>Nano</td>
<td>1.3</td>
<td>Info</td>
<td>3.3</td>
</tr>
<tr>
<td>Info &amp; Cogno</td>
<td>Info</td>
<td>1.3</td>
<td>Cogno</td>
<td>3.5</td>
</tr>
<tr>
<td>Info &amp; Bio</td>
<td>Info</td>
<td>9.4</td>
<td>Bio</td>
<td>4.7</td>
</tr>
<tr>
<td>Cogno &amp; Bio</td>
<td>Cogno</td>
<td>3</td>
<td>Bio</td>
<td>1.2</td>
</tr>
</tbody>
</table>

Shared keywords
The third way of looking at the relation or similarity of two fields here is keyword analysis. Table 6 shows the keyword similarity between pairs of subjects. In the strongest similarity we can see that 483 keywords had a total frequency of 15,855 (28.6%) in Nano articles. These 483 keywords were also used in Bio articles and they accounted for 38 per cent of keyword frequencies in Bio articles. The weakest relation is between Cogno and Bio articles, where
Cogno articles have the smallest number of shared keywords with Bio articles. Figure 3 shows the spanning tree of shared keywords. The number inside nodes are the total number of keywords in the given field.

Table 6. Overlap of title keywords (shared keywords) in each pair of subjects

<table>
<thead>
<tr>
<th>Combinations</th>
<th>No of shared keywords</th>
<th>First field</th>
<th>No of keywords</th>
<th>% of keywords</th>
<th>Second field</th>
<th>No of keywords</th>
<th>% of keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nano &amp; Cogno</td>
<td>25</td>
<td>Nano</td>
<td>755</td>
<td>1.4</td>
<td>Cogno</td>
<td>871</td>
<td>15.2</td>
</tr>
<tr>
<td>Nano &amp; Bio</td>
<td>483</td>
<td>Nano</td>
<td>15,855</td>
<td>28.6</td>
<td>Bio</td>
<td>8,664</td>
<td>38</td>
</tr>
<tr>
<td>Nano &amp; Info</td>
<td>141</td>
<td>Nano</td>
<td>11,526</td>
<td>20.8</td>
<td>Info</td>
<td>4,799</td>
<td>20.4</td>
</tr>
<tr>
<td>Info &amp; Cogno</td>
<td>16</td>
<td>Info</td>
<td>1,216</td>
<td>1.2</td>
<td>Cogno</td>
<td>669</td>
<td>11.7</td>
</tr>
<tr>
<td>Info &amp; Bio</td>
<td>56</td>
<td>Info</td>
<td>1,198</td>
<td>4.7</td>
<td>Bio</td>
<td>1,513</td>
<td>6.6</td>
</tr>
<tr>
<td>Cogno &amp; Bio</td>
<td>68</td>
<td>Cogno</td>
<td>3,006</td>
<td>0.01</td>
<td>Bio</td>
<td>1,241</td>
<td>5.4</td>
</tr>
</tbody>
</table>

Figure 3. Spanning tree for the relations of four subjects based on shared keywords

*Inclusion index* for the similarity of pairs of subjects in terms of shared keywords is presented in Table 7. The highest quantity is the similarity of Bio articles’ keywords (5.6%) to those of Nano articles. The lowest quantity belongs to similarity of Nano articles’ keywords (1.2%) to those of Info articles.

Table 7. Inclusion index for overlap of keywords for pairs of subjects

<table>
<thead>
<tr>
<th>Combinations</th>
<th>First field</th>
<th>Inclusion index (%)</th>
<th>Second field</th>
<th>Inclusion index (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nano &amp; Cogno</td>
<td>Nano</td>
<td>3.3</td>
<td>Cogno</td>
<td>2.8</td>
</tr>
<tr>
<td>Nano &amp; Bio</td>
<td>Nano</td>
<td>3.04</td>
<td>Bio</td>
<td>5.6</td>
</tr>
<tr>
<td>Nano &amp; Info</td>
<td>Nano</td>
<td>1.2</td>
<td>Info</td>
<td>2.9</td>
</tr>
<tr>
<td>Info &amp; Cogno</td>
<td>Info</td>
<td>1.3</td>
<td>Cogno</td>
<td>2.4</td>
</tr>
<tr>
<td>Info &amp; Bio</td>
<td>Info</td>
<td>4.7</td>
<td>Bio</td>
<td>3.7</td>
</tr>
<tr>
<td>Cogno &amp; Bio</td>
<td>Cogno</td>
<td>2.3</td>
<td>Bio</td>
<td>5.5</td>
</tr>
</tbody>
</table>

For keywords, we can also consider how each field imports concepts from or exports concepts to the other fields. Exporting here indicates the percentage of keywords of a given field that have been used by other fields. The percentage of concept export for Cogno articles here is 56
per cent (i.e. 1241+1216+755)/5725*100 = 56%). To calculate import percentage for Cogno articles, the sum of all of its join keywords with the other three fields is divided by the total number of its keywords. The concept import percentage for Cogno articles is 79 per cent. The ratio of import to export (59/79) is 0.7 for Cogno. This ratio indicates that Cogno exports more concepts to the other fields than it imports. Nano seems to have the largest import to export ratio.

Table 8. Import and export of concepts

<table>
<thead>
<tr>
<th>Fields</th>
<th>Importing concepts (%)</th>
<th>Exporting concepts (%)</th>
<th>Ratio of import to export</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nano</td>
<td>50.7</td>
<td>25.8</td>
<td>1.9</td>
</tr>
<tr>
<td>Bio</td>
<td>88</td>
<td>50.1</td>
<td>1.7</td>
</tr>
<tr>
<td>Info</td>
<td>30.8</td>
<td>58.5</td>
<td>0.51</td>
</tr>
<tr>
<td>Cogno</td>
<td>79</td>
<td>56</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Authorship analysis

Authors of one subject field might publish articles in the journals of another subject field. Table 9 shows the number and percentage of authors who published in each pair of subject fields. In the pair of Nano and Bio, 9,478 authors published in the journals of both fields. These authors, given the frequency of their appearance in the articles (counting was not unique counting) were 67.4 per cent (or 50,131 authorships) of Nano articles. The same 9,478 authors accounted for 40.7 per cent or 50,915 authorships in Cogno articles. Figure 4 shows the spanning tree graph of authorship. The number inside nodes are the total number of authors in the given field.

Table 9. Overlap of authors in each pair of subjects

<table>
<thead>
<tr>
<th>Combinations</th>
<th>No of authors publishing in both fields</th>
<th>First</th>
<th>% of authors</th>
<th>Second</th>
<th>% of authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nano &amp; Cogno</td>
<td>1,494</td>
<td>Nano</td>
<td>4,373</td>
<td>Cogno</td>
<td>2,774</td>
</tr>
<tr>
<td>Nano &amp; Bio</td>
<td>9,478</td>
<td>Nano</td>
<td>50,131</td>
<td>Bio</td>
<td>50,915</td>
</tr>
<tr>
<td>Nano &amp; Info</td>
<td>4,130</td>
<td>Nano</td>
<td>28,416</td>
<td>Info</td>
<td>14,434</td>
</tr>
<tr>
<td>Info &amp; Cogno</td>
<td>1,494</td>
<td>Info</td>
<td>9,972</td>
<td>Cogno</td>
<td>4,140</td>
</tr>
<tr>
<td>Info &amp; Bio</td>
<td>5,372</td>
<td>Info</td>
<td>6,262</td>
<td>Bio</td>
<td>10,905</td>
</tr>
<tr>
<td>Cogno &amp; Bio</td>
<td>1,343</td>
<td>Cogno</td>
<td>4,176</td>
<td>Bio</td>
<td>10,535</td>
</tr>
</tbody>
</table>
Table 10 gives the Inclusion index for the similarity of pairs of subjects in terms of shared authors. The smallest number is the similarity of Nano article’s authorship (2.01%) to that of Cogno articles. The largest number is the similarity of Nano articles’ authorship to that of Bio articles.

Table 10. Inclusion index for overlap of authors for pairs of subjects

<table>
<thead>
<tr>
<th>Combinations</th>
<th>First field</th>
<th>Inclusion index (%)</th>
<th>Second field</th>
<th>Inclusion index (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nano &amp; cogno</td>
<td>Nano</td>
<td>2.01</td>
<td>Cogno</td>
<td>8.72</td>
</tr>
<tr>
<td>Nano &amp; bio</td>
<td>Nano</td>
<td>12.75</td>
<td>Bio</td>
<td>6.63</td>
</tr>
<tr>
<td>Nano &amp; info</td>
<td>Nano</td>
<td>5.55</td>
<td>Info</td>
<td>8.92</td>
</tr>
<tr>
<td>Info&amp; cogno</td>
<td>Info</td>
<td>3.22</td>
<td>Cogno</td>
<td>8.72</td>
</tr>
<tr>
<td>Info&amp; bio</td>
<td>Info</td>
<td>11.61</td>
<td>Bio</td>
<td>4.31</td>
</tr>
<tr>
<td>Cogno &amp; bio</td>
<td>Cogno</td>
<td>7.8</td>
<td>Bio</td>
<td>1.07</td>
</tr>
</tbody>
</table>

For authorship, we can also consider inward and outward interactions. Outward interaction is the percentage of authors from a given field that have contributed in the articles of any of the other three fields. Inward interaction is the percentage of authors from any of the other three fields that have contributed in the articles of the given field. Table 11 presents the data. We can see that inward interaction of authors in Nano is 91.6 which was calculated by dividing the sum of contributions of authors of the other fields in Nano articles by the total number of authors of Nano articles. Nano has a larger percentage of outward interactions. The ratio of inward to outward interactions (inward/outward) shows whether the authors of a field contribute to the other fields more than other receiving contribution from the other three fields or it is the opposite. In the case of Nano, 0.82 shows that authors of this field contribute more in the articles of the other three fields than the field receives contribution from the other three fields.

Table 11. Inward and outward activity of authors

<table>
<thead>
<tr>
<th>Fields</th>
<th>Inward(%)</th>
<th>Outward(%)</th>
<th>Ratio of inward to outward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nano</td>
<td>91.6</td>
<td>111.5</td>
<td>0.82</td>
</tr>
<tr>
<td>Bio</td>
<td>48.5</td>
<td>17.2</td>
<td>2.8</td>
</tr>
<tr>
<td>Info</td>
<td>93.9</td>
<td>66.2</td>
<td>1.4</td>
</tr>
<tr>
<td>Cogno</td>
<td>58.2</td>
<td>64.8</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Discussion and conclusions

This study is one of the few studies, if not the only, that have looked at the interdisciplinary relations between the four fields of NBIC (nanotechnology, biotechnology, information technology, and cognitive science). These are known as converging (or convergent) technologies and there is a great deal of hope that their synergy will result in great progress in science. The first level of their convergence is at the science level, which is manifested in interdisciplinary relations. These relations were investigated in this study using a set of different techniques and measures. These combination of techniques and measures together (e.g. inclusion index and spanning tree) have not been used much in the past. The use of different techniques reveals different aspects of interdisciplinary relations. For instance while keyword analysis shows the exchange of concepts between two fields, citation analysis reveals how much...
they have in common in terms of sources of information; and authorship analysis shows something about their human interaction.

The results of the study showed that the highest exchange of citations occurs between Bio and Nano. Bio and Nano also had the strongest link in terms of shared keywords. The highest exchange of authorship was also between Nano and Bio. These results show that Bio and Nano have probably a higher level of convergence and they are more likely to be able to do joint projects. This should not be surprising as there is already a field of research named bionanoscience (Rafols and Meyer, 2010) that indicates the convergence of Nano and Bio. This result is aligned with the findings of past studies (Porter and Youtie, 2009) on high interdisciplinarity of Nano and its convergence with some other fields.

The results also showed that Cogno articles did not have any citations to the journals cited in Bio articles and the two fields of Cogno and Bio did not have any relation in this regard. Cogno also had the weakest link to Nano in terms of authorship as well as author citations. The lowest rate of exchange of concepts also belonged to Cogno and Bio. All of these indicate that cognitive science in the context of Iranian science has not been very successful in establishing relations with the other three fields in NBIC and its interdisciplinary relation with the other three fields are not as strong as it should be. The low integration of Cogno with other fields in this study was not expected for past studies (Bergman et al., 2016) found cognitive science to be very interdisciplinary.

We should note that this study has its own limitations as the data used in the study were only the data of Iranian publications. Interactions between scientific fields in the context of Iranian science is influenced by the way science is managed in Iran. There are two independent and separate ministries in Iran in charge of research. Ministry of Science, Research and Technology manages most areas of science except Medical sciences that are managed by the Ministry of Health. This might create some hindrances for the relations of certain subfield of cognitive science (such as certain areas of neuroscience that falls under the administration of the Ministry of Health) with other areas (i.e. Nano, Bio and Info). This governance aspect of science might explain the weak relationship of Cogno with the other fields in this study. It will be interesting to do the same kind of analyses on the global data to see if the relations in Iranian NBIC are different from those in the international science. The other issue is that most of the analyses in this study, similar to many other bibliometrics studies, relied on the subject categorizations of journals in Scopus database. In spite of these limitations, use of different techniques and measures helped create a better picture of the interdisciplinary relations between NBIC. Spanning tree could show the weakest and strongest connections and inclusion index showed the similarity. It is recommended that future studies use a combination of techniques for the study of interdisciplinary relations. The study also focused only on publications, and there is a need to investigate the convergence of NBIC at the technology (e.g. patent), market, and industry levels.

References


731
Mapping the development trajectory of 3D printing technologies

Lili Wang¹ and Shan Jiang²

¹ wang@merit.unu.edu
UNU-MERIT, Maastricht (The Netherlands)

² jiangs@whlib.ac.cn
Wuhan Library, Chinese Academy of Sciences (China)

Abstract

Using patent and citation analysis, this study explores the development trajectory of 3 dimensional printing (3DP) technologies. Different from the existing literature, we disentangle non-3DP technologies from those in the 3DP domain. Our results show that 3D printing technologies invented by the U.S. have played in a dominating role in the main path. In particular, new branches of various types of 3D technologies have all developed from American patents. There was only one main path in 3D printing technologies during 1998 and the early 2000s. However, great diversification can be observed after the mid-2000s. China has been catching up after 2012, in technologies related to both printing materials and printing machines. By tracking technology domains in the patents cited by 3DP technologies, we find that, in the period between 2004 and 2012, technologies from non-3DP domains contributed greatly in shaping the development of 3D printing materials and printing machines. At a disaggregated level, this study provides insights on how technologies from various fields contributed to different 3D printing technology clusters. The methods carried out in this study can be also applied to exploring technological developments and knowledge flows in other fields.

Conference Topic
Patent analysis
Citation and co-citation analysis

Introduction

It is widely believed that knowledge flows can be tracked by patents and citations (Jaffé and Trajtenberg, 2005; Verspagen 2007; Hummon and Dereian 1989). Citations in one patent reveal that the knowledge in this patent has been developed from (or is related to) the knowledge described in other patents published earlier. By tracing the inventor’s name, country and year in which the patents were granted, one can also view the locations where the technological innovation took place and the year when the patent was publicly announced. There are clusters or sub-streams in the development of technological trajectory. The main path, however, provides
the most important information on the major stream of knowledge flows and represents the greatest connectivity of different technologies (Verspagen 2007; Hummon and Dereian 1989).

Three-dimensional (3D) printing, known as additive manufacturing, is a process of creating physical objects from a digital design. In the digital era, 3D printing is regarded as socially transformative technology with great socioeconomic implications (Ratto and Ree 2012). Several studies have been conducted, using citation analysis, to map the development of 3D technologies (Huang et al. 2017; Kim et al. 2016). The existing literature has examined 3D printing trajectory with citation links without disentangling technological domains (i.e. citation resources). We argue that understanding the technological contribution from other technological domains is crucial for a deeper understanding of the emergence of new technologies. To this end, our study covers two types of knowledge sources. On the one hand, we investigate the contribution of different countries in the main path of 3D printing development. This reflects the knowledge flows across geographical regions. On the other hand, we disentangle the contribution of different technological domains, including the 3D printing field itself and the domains of non-3D printing technologies.

Data and Methodology

Patent data have been collected through the Thomson Innovation (TI) platform from Derwent World Patent Index (DWPI)\(^1\). 3D printing patents were extracted based on a series of keywords, including 3D print, Additive Manufacturing, Rapid prototyping, and the key 3D printing techniques such as Stereolithography and Fused Deposition Model. The query we used is: TI=((3d) or (three ADJ dimension*) or (3 ADJ dimension*)) adj (print* or manufact*)) OR AB=((3d) or (three ADJ dimension*) or (3 ADJ dimension*)) adj (print* or manufact*)) OR TI=((Rapid ADJ protyp*) or (Rapid ADJ manufacturing) or (Additive ADJ manufacturing) or bioprint* or Stereolithograph* or (Fused ADJ Deposition ADJ Model*) or (Laser ADJ Sinter*) or (Direct ADJ Metal ADJ Deposition) or (Layered ADJ Object ADJ Manufact*) or (Select* ADJ laser ADJ melt*) or (Electron ADJ beam ADJ melt*)) OR AB=((Rapid ADJ protyp*) or (Rapid ADJ manufacturing) or (Additive ADJ manufacturing) or bioprint* or Stereolithograph* or (Fused ADJ Deposition ADJ Model*) or (Laser ADJ Sinter*) or (Direct ADJ Metal ADJ Deposition) or (Layered ADJ Object ADJ Manufact*) or (Select* ADJ laser ADJ melt*) or (Electron ADJ beam ADJ melt*))\(^2\).

Based on this query, we extracted in total 37,803 patents, which were further distributed into 19,961 patent families. The dataset covers all available patents from any year and any patent authorities covered by DWPI.


\(^2\) The final data collection was done on 7 March 2017.
Citations were gathered from Derwent Patents Citations Index (DPCI). For all the 3D printing patents collected above, we have in total 217,109 patent citations, which are from 44,664 patent families.

To track knowledge flows according to time and location, we label year and country for each patent family. The year is defined by the earliest publication year of all the patents in the same patent family. Country label is defined by the nation of inventor(s). If there are multiple inventors in one patent publication, we label only the country of the first inventor. In creating the main technological path of technological development in the long run, we follow the method proposed by Liu et al.(2013). The technological trajectory is mapped with software Pajek.

Among the cited 44,664 patent families, this study distinguishes between 3DP and non-3DP patents. The fields of non-3DP patents are examined based on the DWPI Manual Code, which was assigned by teams of Thomson Reuters analysts who have been specially trained in the application of these codes (Larner 2013). We first collect the DWPI codes at the third level (e.g. A11-A or A12-C), following which all the DWPI manual codes were aggregated at the second level (e.g. A11 or A12).

Results and discussions

Main path of 3D printing (3DP) technologies

Based on the citation linkages, the main path of 3D printing technologies is captured by Figure 1. According to the relatedness of technologies, the trajectory of 3D printing technologies is further classified into five technological clusters. The first one, shown in red squares, presents technologies related to Additive manufacturing and Stereolithography. The second cluster, shown in blue squares, includes patents that mainly introduced methods to derive data for forming high resolution three-dimensional object. The third one, in green, indicates a group of patents applying technologies such as beam Stereolithography equipment and Optical molding apparatus etc. The fourth group, in purple, consists of technologies related to 3D printing materials. And the last group, in brown, is comprised of technologies related to 3D printers.
Figure 1. Technological trajectory map of 3DP technologies
The earliest 3D printing technology is represented by EP171069A2_1985_US, which was filed in 1984. In this patent family, according to the application year, the earliest patent was US4575330A filed in 1984, which proposed a type of technology named Stereolithography. In 1986, the inventor of this patent, Hull Charles, co-funded 3D systems\(^3\) – the first 3D printing company in the world.

In the trajectory map, the arrow line illustrates citation relations between patent families, while the arrow head is directed from the cited to the citing ones. In theory, the year of citing patent families should be later than the cited ones. However, due to the fact that a patent family includes a series of patents with different publication dates, and that the year labeled on the map is defined only by the earliest publication year of all the patents in the same patent family, it can happen that a patent family seemingly cited one from an even later publication year. For instance, the earliest patent (A1) in patent family was published in 2006. Thus we labeled the publication year of patent family A as 2006. In patent family B, the earliest patent was published in 2004 and we labeled the publication year of patent family B as 2004. However, there is one patent (B2) published in 2007 that cited A1. Thus at the patent family level, we will have patent family A (labeled in the year of 2006) that was cited by patent family B (labeled in the year of 2004). Figure 2 illustrates the misalignment citations between patent families.

![Figure 2. Example of misalignment citations between patent families](image)

Apart from the few exceptional cases with misalignment citations\(^4\), the rest citation links in the main path are all with logical years. The main path (Fig.1) shows that, after 1998, 3D printing technologies formed one main stream which kept a stable path for the following two decades. However, forked technological streams are observed after 2004. Printing materials and printing machines are the two main clusters developed simultaneously in the later stage. Within either cluster, there are also sub-branches with special technological focus.

\(^3\) [https://www.3dsystems.com/our-story](https://www.3dsystems.com/our-story).

The technological path is dominated by the U.S., with an incidental presence of Japan (JP) and Germany (DE). China (CN) appeared on the map only after 2012, in both technological clusters dealing with printing materials (in purple squares) and printing machines (in brown squares). South Korea (KR) showed up also relatively late on the path, in particular in the track of printing machine related technologies (in the end of brown branch).

**Technological resources (3D printing vs. non 3D printing technologies)**

![Graph showing the share of non 3D printing citations (1986-2015)](image)

*Figure 3. Share of non 3D printing citations (1986-2015)*

Note: The year represents the publication year of the earliest patent in one 3D printing patent family.

Figure 3 shows the share of citation from non-3DP patents. After the introduction of the earliest 3DP patents in the 1980s, the share of non-3DP patent citations decreased over time till the early 2000s. This shows that, with the increasing number of 3DP technologies, more and more knowledge from this domain can be used to further develop new technologies in the following years, at least till 2003. Thus the share of citations from other technology domains remained lower. The lowest share was reached in 2003, which was around 70 per cent. This also indicates that in 2003, on average, 3DP inventions were developed 30% based on 3DP domain itself and 70% on knowledge outside. Afterwards the share of non 3DP citation increased, in particular in the period from 2004 to 2012. This interesting evidence suggests that, in spite of the growing of pool of 3DP technologies, 3DP inventors applied more extra knowledge from other technology domains after 2003. This signals that 3DP technology entered a new era after 2003, with new 3DP breakthroughs influenced by other technology domains.
Technological contribution from non-3DP

Following the trajectory map and the five technological clusters, this section analyses the contribution of non-3DP in the development of 3D printing technologies.

For the top non-3DP patents cited by the five clusters in the main path, all the assigned Derwent manual codes were collected and aggregated at the third level (e.g. A11 or A12). Table 1 provides the major non-3DP fields that have contributed to the five technology clusters.

<table>
<thead>
<tr>
<th>3D printing Clusters</th>
<th>DWPI Manual Codes</th>
<th>Code details of non-3DP technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>T01</td>
<td>DIGITAL COMPUTERS</td>
</tr>
<tr>
<td></td>
<td>A11</td>
<td>PROCESSING POLYMERS INCLUDING EQUIPMENT</td>
</tr>
<tr>
<td></td>
<td>M22</td>
<td>CASTING, POWDER METALLURGY</td>
</tr>
<tr>
<td></td>
<td>L01</td>
<td>GLASS, VITREOUS ENAMELS</td>
</tr>
<tr>
<td></td>
<td>M13</td>
<td>NON-ELECTROLYTIC COATING</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>A11</td>
<td>PROCESSING POLYMERS INCLUDING EQUIPMENT</td>
</tr>
<tr>
<td></td>
<td>A12</td>
<td>POLYMER APPLICATIONS</td>
</tr>
<tr>
<td></td>
<td>G06</td>
<td>PHOTOGRAPHIC MATERIALS AND PROCESSES</td>
</tr>
<tr>
<td></td>
<td>T01</td>
<td>DIGITAL COMPUTERS</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>A05</td>
<td>CONDENSATION POLYMERS</td>
</tr>
<tr>
<td></td>
<td>A08</td>
<td>ADDITIVES</td>
</tr>
<tr>
<td></td>
<td>A11</td>
<td>PROCESSING POLYMERS INCLUDING EQUIPMENT</td>
</tr>
<tr>
<td></td>
<td>A12</td>
<td>POLYMER APPLICATIONS</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>A11</td>
<td>PROCESSING POLYMERS INCLUDING EQUIPMENT</td>
</tr>
<tr>
<td></td>
<td>A12</td>
<td>POLYMER APPLICATIONS</td>
</tr>
<tr>
<td></td>
<td>L02</td>
<td>REFRACTORIES, CERAMICS, CEMENT</td>
</tr>
<tr>
<td></td>
<td>T01</td>
<td>DIGITAL COMPUTERS</td>
</tr>
<tr>
<td></td>
<td>X25</td>
<td>INDUSTRIAL ELECTRICAL EQUIPMENT</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>S06</td>
<td>ELECTROPHOTOGRAPHY AND PHOTOGRAPHY</td>
</tr>
<tr>
<td></td>
<td>T04</td>
<td>COMPUTER PERIPHERAL EQUIPMENT</td>
</tr>
<tr>
<td></td>
<td>W01</td>
<td>TELEPHONE AND DATA TRANSMISSION SYSTEMS</td>
</tr>
</tbody>
</table>

Note: Summarized from the top non-3DP patents cited by the 3DP patents in five clusters.

In the development of 3D printing technologies, processing polymer (including equipment and applications) played a crucial role. The polymer-related processing technology, equipment, materials applications (plastic, resin, etc.) contributed to all the 3DP clusters except cluster 5.
Figure 4. Contribution of non-3DP technologies to five 3DP clusters

Knowledge flows between top non-3DP technologies to the five clusters are summarized in Figure 4. In cluster 1, the major contribution of non-3D printing technologies came from the field of digital computer processing techniques (3-dimensional image modelling), metal powder, and metal laser coating. In the second technology cluster, photosensitive material technologies played an important role in promoting the development of 3D printing. In Cluster 3, in which the earliest patents of the 3DP patent families were published mainly between 2000 and 2006, the major non-3DP technological contribution originated from the polymer field, such as condensation polymer and polymer addition. Different from Cluster 2, the non-3DP patents cited in Cluster 3 are mainly related to the polymer materials rather than polymer equipment. In cluster 4, the non-3DP citations were from various fields, including polymer processing applications, ceramic materials, computer technology, industrial equipment and so on. This facilitates the technology diversity of 3D printing technologies in this group. In cluster 5, non-3DP technologies related to image technology, computer peripherals, inkjet printers and digital transmission systems began to affect 3D printing technology.

Conclusions

Towards understanding the development of 3D printing technologies, this study maps the trajectory of 3D printing patents and explores the contribution of various knowledge sources. Different from previous studies, this paper tracks knowledge sources in contributing to the development of 3D printing technologies. On the one hand, we illustrate the contribution of different countries in the main path of 3D printing development. On the other hand, we disentangle the contribution of non-3DP technologies from that of 3DP domain.
The linkages between patent families from different countries and different years help understand the role major nations played in various stages. The trajectory map shows that the 3D printing technological path has been dominated by the U.S., with an incidental presence of Japan and Germany. After 2012, however, China witnessed a sharp rise in technologies related to 3D printing materials and 3D printing machines.

By differentiating knowledge sources from various domains, our results suggest that 3D printing technology entered into a new era after 2003. In the period of 2004 and 2012, technologies from non 3DP domains contributed greatly in shaping the development of 3D printing materials and 3D printing machines. Based on the classification of non-3D print citations, this study also provides insights on how technologies from various fields contributed to different 3D printing technology clusters.

It is noteworthy that this study defines countries of patents based on the first inventors. Thus a collaborated patent is assigned only to one country. Admittedly, various collaborations happen across the global innovation value chain. Testing such knowledge flows related to co-inventorship may be an interesting subject for future research.

References


740
Knowledge Spillovers among Semiconductor Companies with Different Technology Positions and with Different Roles in the Industry Chain

Chun-Chieh Wang¹  Mu-Hsuan Huang²

¹wangcc@ntu.edu.tw
Department of Bio-Industry Communication and Development, National Taiwan University, Taipei, Taiwan

²Corresponding Author: mhhuang@ntu.edu.tw
Department of Library and Information Science, National Taiwan University, Taipei, Taiwan

Abstract
This study focuses on the knowledge spillovers in the semiconductor industry by employing patent bibliometrics and social network analysis methods. Based on the technological development of wafer diameters, the analysis is divided into three time periods: 6-inch (1976-1991), 8-inch (1989-1999) and 12-inch (1997-2011). Two types of knowledge spillover relationships between semiconductor companies are examined, of which strong ties are defined as the R&D cooperation activities between companies, while weak ties are represented by the patent citations between semiconductor companies. We apply two divisions to categorise semiconductor companies in this study. First, based on each company’s technology position in the knowledge spillovers network, companies are classified as technology leaders, technology brokers, technology followers, or technology isolated companies. Second, based on the role of each firm in the industry chain, semiconductor companies are divided into six types, which are—integrated device manufacturer (IDM), IC design, foundry, packaging & testing, IC equipment, and others. Study results indicate that technology leaders are the main sources of semiconductor knowledge spillovers among the four distinct technology positions. Furthermore, IDM and foundries are the sources of semiconductor knowledge spillovers among all roles within the industry chain. Consequently, we suggest that monitoring the development of technology leaders, such as IDMs and foundries, is important in formulating strategies for the research and development of the semiconductor industry. In addition, it is suggested that foundries transform their relationships with IDMs from vertical outsourcing manufacturing to horizontal R&D cooperation.

Keywords
Patentometrics
knowledge spillover
semiconductor industry

Conference Topic
Patent analysis

Introduction
To improve understanding of the characteristics of knowledge spillovers, it would be beneficial to systematically characterise companies when conducting research in this area. Andersen (2006) analyzed the characteristics of enterprise knowledge spillovers in European and non-European countries by positioning the technology of companies. Moreover, Podolny, Stuart and Hannan (1996) divided companies into technology leaders, technology brokers, technology followers and technology isolated companies through the calculation of two indices—technological crowding and technological prestige, which are based on the patent citations between semiconductor companies. Also, Podolny et al. (1996) inspected how the technology positions of companies affect their survival rates. Similar applications that classify companies based on their technology positions can be used to predict the formation of technology strategic alliances among companies with different technology positioning (Stuart, 1998), and be used to evaluate the shares of different technology positions in distinct industries across different countries (Andersen, 2006; Okamura & Vonortas, 2006). Apart from the classification of technology positions, many industries are categorised by their roles in the industry chain. For
instance, the semiconductor industry divides their business activities into integrated device manufacturer (IDM), IC design companies, foundries, packaging & testing companies, IC equipment companies and others (such as research institutions, terminal products manufacturing and selling companies).

The aim of this study is to discuss the channels of knowledge spillovers, considering R&D cooperation as strong ties of knowledge spillovers between companies and the patent citation connections as weak ties of knowledge spillovers (Wang, Sung, Chen & Huang, 2017). By the means of positioning companies into technology leaders, technology brokers, technology followers, and technology isolated companies, as well as categorising companies into integrated device manufacturer (IDM), IC design, foundry, packaging & testing, IC equipment and others based on their roles in the industry chain, we examine the differences in the performance of knowledge spillovers between the distinct types of companies and further analyze the degree of involvement and the receivers of knowledge spillovers.

Koka, Madhavan, and Prescott (2006) stated that two conditions affected the formation of enterprise networks, which include the uncertainty of the environment and the supply of resources. In order to fully explore the evolution of company networks within an industry, it is necessary to look for an industry that has developed for a considerable period of time and that has comprehensive data resources. Since the semiconductor industry meets the criteria mentioned above, it was selected for analysis in this study. In this paper, we analyze the knowledge spillover network of semiconductor companies over 36 years, divided into three time periods which correspond to the development of the 6-inch, 8-inch and 12-inch wafer diameters. By means of patent bibliometrics, social network analysis and statistical tests, we examine the differences between the various types of semiconductor companies as exhibited by their strong ties and weak ties of knowledge spillovers.

Knowledge Spillover Network

**Technology positions**

The relationship of patent citations between companies can be summarised and represented by two indices—technological crowding and technological prestige. We apply these indices and further divide the technology position of companies into four categories, that is, technology leaders, technology brokers, technology followers and technology isolated companies. Podolny et al. (1996) set technological crowding as the x-axis and technological prestige as the y-axis. Companies located in the first quadrant were defined as technology leaders, who have both high technological crowding and high technological prestige. Technology leaders focus on the research and development of mature technology while maintaining leading positions in the industry. Moreover, technology leaders have been considered the originators of knowledge spillovers in the semiconductor industry. Companies situated in the second quadrant represent technological brokers. Technology brokers have low technological crowding but high technological prestige. These companies tend to focus on the development of innovative and unexplored technology areas. The region in the third quadrant indicates that a certain firm is a technological isolated company, which has low technological crowding as well as low technological prestige. Technological isolated companies develop technologies that are rather out of the ordinary, and they either focus on cutting-edge technology that has not yet been adopted by the industry, or out-of-date technology that is soon to be obsolete. Companies in the fourth quadrant are defined as technology followers, who have high technological crowding but low technological prestige. Technology followers focus on the development of mature technology but rarely have innovative contribution to the industry. Related studies on the citation networks of companies classified according to their technology positions particularly focus on how the different technology positions affect a company’s survival rate (Podolny et
al., 1996); how a company’s technology position predicts its formation of technology strategic alliances (Stuart, 1998); and the distribution of countries based on their technology positions, using data from different industries (Andersen, 2006; Okamura & Vonortas, 2006).

Roles in the Industry Chain
According to the definition of companies in a semiconductor industry chain, semiconductor companies can be divided into integrated device manufacturer (IDM), IC design, foundry, packaging & testing, IC equipment and others (such as research institutions and terminal products manufacturing and selling companies). This paper classifies companies in the semiconductor industry into six roles by referring to the introduction in the companies’ official websites, Wikipedia and the Semiconductor Industry Yearbook published between 1999 and 2011. Each company can only be classified under one role with no overlapping, and the classification results were confirmed by industry experts.

Methodology
Data Collection
In this study, patents granted by the USPTO between 1976 and 2011 were gathered for analysis and data related to R&D cooperation was collected from the Semiconductor Industry Yearbook published between 1999 and 2011. Here we limit our patent citations and data sources to the United States Patent Classification (USPC) System, which is maintained by the U.S. Patent and Trademark Office (USPTO). Patents related to semiconductor technologies are categorised in the USPC under Electronic-Semiconductor Devices and its patent classes are 257, 326, 438 and 505 (Hall, Jaffe, & Trajtenberg, 2001). In addition, the R&D cooperation records between semiconductor companies were obtained and compiled from the Semiconductor Industry Yearbook, which is published by the Industrial Economics & Knowledge Center (IEK) of the Industrial Technology Research Institute in Taiwan. The yearbook has 21 volumes from 1999 up to 2011, and documents the development trajectories and issues in the semiconductor industry every year in Taiwan and other major countries. It also presents the dynamics and major changes in the industry, providing authoritative references to R&D cooperation activities in the semiconductor industry.

Patent Bibliometrics
Patent bibliometrics (or patentometrics) is a theoretical method using mathematics, statistics and logic. It studies and analyzes the quantity, quality and application of patent literature, e.g. patent counts and patent citations (Narin, 1994). Patent bibliometrics can be used to understand the development of patented technologies (including that of individuals, organisations and countries). What is more, it can be used to study the links between researchers, organisations and countries through the relationship of patent citations.

Patent analysis calculates the patent counts and the frequency distribution of patents based on the selected units of analysis (e.g. country, company/organisation, inventor and technology field), and can be used to identify the major activities of the selected units. Patent citation analysis focuses on the references (including patent and non-patent references) cited in the patent specification. As a result, potential links can be explored through patent citation counts and citation relations. Patent citations are similar, but not equal to citations in literature. The inventor/applicant might cite documents, but examiners also cite documents in their novelty reports. This study mainly focuses on the evaluation of weak ties of knowledge spillovers in the semiconductor industry using patent bibliometrics. In order to distinguish the weak ties network during the three wafer diameters periods, here we apply the Activity Index as the indicator to screen and select representative patent citation links.
Activity Index

The Activity Index (AI) is an indicator that was originally defined as the share of a given technological field among all patents held by a company, divided by the share of the given technological field among all U.S. patents. This indicator also showed the relative activeness of a company in different technological fields in order to identify which technological fields the company specialised in (Pavitt and Soete, 1980). In this study, we adjusted the indicator and applied it to measure the activeness of citing connections from one company to another. The AI is the ratio between the percentage of the patent citations to a cited company among all outbound citations of a citing company and the percentage of the patent citations to the cited company among all outbound citations of all companies. It is worth noting that the activeness, in this case, is not ‘absolute’ but ‘relative’. ‘Relative’ means that the company is more active compared with the given technological field, which does not indicate that the company has more absolute technical advantages than the other companies. The AI value ranges from 0 to $+\infty$. It indicates whether the company is relatively less active (AI < 1) or relatively more active (AI > 1) to the given field. The formula for the AI is shown as follows:

\[
\text{Activity Index (AI)} = \frac{\text{patent citation counts from the citing company to the cited company}}{\text{total patent citations from all companies to the cited company}} / \frac{\text{total outbound patent citations of the citing company}}{\text{total outbound patent citations of all companies}}
\]

Data Sampling

This study draws upon patent citation relations to analyze the knowledge spillovers between companies. However, the number of patent citations received by a company varies relative to its number of patents held, in which the likelihood of being cited increases for companies with more patents. In terms of patent citation counts, should one single citation count represent the weak ties of knowledge spillover between two companies? In order to select representative patent citation relations between companies, this study adopts not only citation counts but also the Activity Index for data sampling, explained as follows:

- **Condition 1**: The number of patent citations from the citing company to the cited company must be higher than the median of patent citations from the citing company to all companies.
- **Condition 2**: The ratio between the patent citation count from the citing company to the cited company and the total outbound citations of the citing company must be higher than the ratio between the patent citations from all companies to the cited company and the total outbound citations of all companies in the industry. In other words, the AI value of the link from one company to another must be higher than 1.

In the sampling process, patent citation relations that meet condition 1 were regarded as high ‘absolute’ citation relations of the citing companies. Those that meet condition 2 were regarded as high ‘relative’ citation relations. Therefore, sampled patent citation links that meet both conditions 1 and 2 are considered representative for measuring the knowledge spillovers between companies in this study.

Technology Positions and the Roles in the Industry Chain of Semiconductor Companies

In order to investigate the differences of strong ties and weak ties of knowledge spillovers based on the characteristics of companies in semiconductor industry, we distinguish the technology position of companies. The methodology used in this research is adopted from Podolny et al. (1996), who applied technological crowding and technological prestige to divide companies into four types of technology positions—technology leader, technology broker, technology follower and technology isolated company.
Technological Crowding

Technological crowding evaluates the niche overlap of R&D cooperation of each company, the more overlapping of R&D activities the higher the similarities in innovative capabilities between a particular company and other companies in the field. Podolny et al. (1996) applied patent citations to calculate technological crowding, while Breschi, Cassi, and Malerba (2006) extended the calculation of technological crowding to the level of organizations. The formula is defined as

\[ A_i = \sum_{j \neq i} \frac{C_{ij}}{C_i} \]

where \( C_i \) is the patent citation counts of company \( i \), \( C_{ij} \) is the number of shared references of company \( i \) and company \( j \) (i.e. bibliographic coupling). \( A_i \) therefore represents the degree of technological overlap of company \( i \) with the other companies.

Technological Prestige

Bonacich (1987) proposed a method to measure technological prestige using the power centrality index. It considers the degree centrality based on the direct paths to a company as well as the degree centrality based on indirect paths (i.e. the centrality of the neighboring companies). This shows that the advantages that a company has in the network not only depends on the centrality based on direct links, but also the centrality based on indirect links.

In social network analysis, if some of the nodes have a higher number of connections than the others, they will be considered as having dominant positions in the network due to having a greater sphere of influence. However, nodes with the same number of connections do not imply that they have the same level of influence in the network. Bonacich (1987) pointed out that the calculation of the Power Centrality Index considers both paths linked to a target node and paths that are linked to the neighbouring nodes of the target node. In other words, the calculation of power centrality is formed by the direct paths and the indirect paths of a company. For instance, assessing the power centrality of a particular node considers both the first layer of nodes linked to it and the surrounding nodes that are linked to the nodes in the first layer. Nodes in the first layer are affected by the second layer of nodes linked to them, hence, the power centrality of the central node is calculated based on all the nodes that can be linked to the central node and the relations between the nodes. The formula of power centrality is defined as

\[ C_i(\alpha, \beta) = \sum_j (\alpha + \beta C_j)R_{ij} \]

where \( R_{ij} \) is the relation matrix, representing the relations between nodes. \( \alpha \) is the number of connections with surrounding nodes. \( \beta \) is the attenuation factor and has a value between -1 and 1.

Technology Positions of Companies

Through the calculation of technological crowding and technological prestige, we further divide semiconductor companies into four categories, which are:

- Technology leaders: Companies in the first quadrant have both high technological crowding and high technological prestige. Technology leaders focus on the research and development of mature technology and maintain leading positions in the industry. These companies are often considered as the sources of knowledge spillovers in the industry. IBM, TI and Intel are examples that are considered as technology leader in the semiconductor industry in this study.

- Technology brokers: Companies in the second quadrant have low technological crowding but high technological prestige. These companies tend to focus on the development of innovative and unexplored fields. For instance, Semiconductor Energy
Lab., NEC and Philips are representatives of technology brokers in the semiconductor industry in this study.

- Technology followers: Companies in the fourth quadrant have high technological crowding but low technological prestige. Technology followers focus on the development of mature technology but rarely have innovative contributions to the industry. Cypress Semi, VLSI Technology and Integrated Device Technology are typical technological followers in the semiconductor industry.

- Technology isolated companies: Companies in the third quadrant have both low technological crowding and low technological prestige. Their development area is out of the ordinary, and they either focus on advanced technology that has not been yet been adopted by the industry, or out-of-date technology that is no longer being used, Eastman Kodak and Analog Devices represent technology isolated companies in this study.

Roles of Companies in the Industry Chain

In order to investigate knowledge spillovers in the semiconductor industry at the organization level, we further group companies by their role in the semiconductor industry chain. This way, we can examine the differences between a company’s strong ties and weak ties of knowledge spillovers based on their different roles in the industry chain.

Five common roles of companies in the semiconductor industry can be elaborated as follows:

1. Integrated Device Manufacturer (IDM): These companies are capable of vertical integration. That is to say, IDMs have capabilities in IC design, manufacturing, packaging, testing and selling of wafers. In addition, some of the IDMs release residual wafer to IC design companies, creating the IDM-Foundry business model. For instance, Micron Technology, Samsung and IBM are IDM companies.

2. IC design: These companies only design and sell wafers and do not have equipment or utilities for wafer manufacturing. They can be regarded as fabless companies in the semiconductor supply chain. They outsource packaging, testing and other work to professional wafer manufacturers and packaging & testing companies. Semiconductor Energy Lab., Sun Microsystems and IMEC are IC design companies examined in this study.

3. Foundry: These companies have professional equipment and technologies in wafer production and they mainly produce wafers for their clients. TSMC, UMC and GlobalFoundries are representatives of foundries.

4. Packaging & testing: These companies possess equipment for wafer packaging and they primarily package and test a variety of wafers for their clients. Packaging is an activity that leads the function signal on the chip to external circuits through a carrier and provides protection against impact and corrosion. There are two stages of testing—before and after packaging. It is essential to test chips to verify the quality of the product. ASE, STATS ChipPAC and Amkor are semiconductor packing & testing companies discussed in this study.

5. IC Equipment: Providers of equipment and devices for semiconductor manufacturing, chip packaging and testing. Applied Materials, Shin-Etsu Chemical and Agilent Technologies represent IC equipment companies in the semiconductor industry chain.

Results

The Involvement of Knowledge Spillovers based on the Different Technology Positions and the Different Roles of Companies in the Industry Chain

First, we observe the shares of the different technology positions among the companies and further explore the differences based on their positions in the knowledge spillovers network. When a company’s position is more central than the others, it has a higher degree of knowledge...
spillovers in the network and more knowledge spillover relations when compared to other companies. Table 1 is the proportion and number of companies based on the different technology positions, divided by three periods. The total number of companies during the 6-inch, 8-inch and 12-inch wafer periods are 129, 194 and 292, respectively. Examining the proportion of companies according to their technology positions, technology isolated companies count for more than 50% across all three periods. Technology followers come next at around 27% and increase to 35% during the 12-inch wafer period. Technology leaders only account for 13% and decrease to less than 10% in the 12-inch wafer period. Last, technology brokers account for the least share, at only 1% during the 6-inch and 12-inch wafer periods and 6% in the 8-inch wafer period.

Second, we observe the proportion of companies based on their different roles in the industry chain. This is followed by an investigation of the differences in the knowledge spillovers networks based on the roles of companies in the industry chain. Table 2 presents the proportion and number of companies based on the six roles in the industry chain across three periods. Companies that are classified as “Others” account for the largest share, which comprise 29% of all companies during the 6-inch and 8-inch wafer periods and rise to 32% in the 12-inch wafer period. IDMs are the second largest group, which account for 37% during the 6-inch wafer period and fall to 22% during the 12-inch wafer period. IC design companies account for 13% in the 6-inch wafer period and slightly rise to 21% during the 12-inch wafer period. The proportions of foundries and packaging & testing companies slightly increased from the 6-inch to the 12-inch wafer period, but neither type of company comprises over 10% of the total share. IC Equipment companies show a decreasing trend from the first to the last period, accounting for 7% to 5% of all semiconductor companies.

<table>
<thead>
<tr>
<th>Technology positions</th>
<th>6-inch wafer</th>
<th>8-inch wafer</th>
<th>12-inch wafer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology Leaders</td>
<td>17(13.18%)</td>
<td>16(8.25%)</td>
<td>29(9.93%)</td>
</tr>
<tr>
<td>Technology Brokers</td>
<td>2(1.55%)</td>
<td>13(6.70%)</td>
<td>4(1.37%)</td>
</tr>
<tr>
<td>Technology Followers</td>
<td>35(27.13%)</td>
<td>54(27.84%)</td>
<td>104(35.62%)</td>
</tr>
<tr>
<td>Technology Isolated</td>
<td>75(58.14%)</td>
<td>111(57.22%)</td>
<td>155(53.08%)</td>
</tr>
<tr>
<td>Total</td>
<td>129(100%)</td>
<td>194(100%)</td>
<td>292(100%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6-inch wafer</th>
<th>8-inch wafer</th>
<th>12-inch wafer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Companies</td>
<td>Patents</td>
<td>Companies</td>
</tr>
<tr>
<td>IDM</td>
<td>(37.98%)</td>
<td>(83.49%)</td>
</tr>
<tr>
<td>IC Design</td>
<td>(13.95%)</td>
<td>(2.76%)</td>
</tr>
<tr>
<td>Foundry</td>
<td>(6.98%)</td>
<td>(0.03%)</td>
</tr>
<tr>
<td>Packaging &amp;</td>
<td>(3.88%)</td>
<td>(0.57%)</td>
</tr>
<tr>
<td>Testing</td>
<td>(7.75%)</td>
<td>(2.37%)</td>
</tr>
<tr>
<td>Others</td>
<td>(29.46%)</td>
<td>(10.78%)</td>
</tr>
<tr>
<td>Total</td>
<td>129(100%)</td>
<td>12,724</td>
</tr>
</tbody>
</table>
Knowledge Spillovers from Companies with Different Technology Positions over Time

By examining the network diagrams during the three periods, we can examine the performance of knowledge spillovers of companies based on their different technology positions. Technology leaders gradually moved to central positions in the network from the 6-inch to the 12-inch wafer period. By contrast, the network positions of technology brokers have decentralized over the three periods. As for technology followers and technology isolated companies, only a few companies occasionally move close to the center of the network.

We conduct an ANOVA to test the network centrality values of companies grouped by technology position over three periods. Table 3 shows the results of a multiple comparison analysis of network centrality between companies with different technological positions; only significant post-hoc test results are presented. In both strong ties and weak ties of knowledge spillover networks, the involvement of technology leaders are significantly higher than both technology followers and technology isolated companies. In other words, technology leading companies connect to more companies through the strong ties and weak ties of knowledge spillovers during all three periods. Secondly, technology brokers have significantly higher involvement and more connections through strong ties and weak ties of knowledge spillovers than technology followers or technology isolated companies during the 8-inch wafer period.

Table 3 ANOVA Test of the Network Centrality of Companies based on Different Technology Positions

<table>
<thead>
<tr>
<th>Type of knowledge spillover network</th>
<th>6-inch wafer</th>
<th>8-inch wafer</th>
<th>12-inch wafer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong Ties of Knowledge Spillovers</td>
<td>L&gt;F, I</td>
<td>L&gt;F, I</td>
<td>L&gt;B, F, I</td>
</tr>
<tr>
<td>Weak Ties of Knowledge Spillovers</td>
<td>L&gt;F&gt;I</td>
<td>L&gt;F, I</td>
<td>L&gt;F&gt;I</td>
</tr>
</tbody>
</table>

Note: L=Technology Leaders, B=Technology brokers, F=Technology followers, I=Technology Isolated Companies. The mean difference is significant at the .05 level.

Table 4 shows the results of a multiple comparison analysis testing the average number of connections between one type of company and another based on technological position; only significant post hoc test results are presented. Between the four types of companies, there are a total of 10 possible pairings in the undirected strong ties network and a total of 16 pairings in the directed weak ties network. Thus, we can examine which two types of companies tend to connect to each other in the knowledge spillover networks. The results indicate that all four types of companies tend to make more connections with technology leaders through both strong ties and weak ties in the network, while a significantly lesser number of connections exist between their pairings with the other three types of companies. Technology brokers also make significantly more connections with other companies, especially when compared to technology followers or technology isolated companies during the 6-inch and 8-inch wafer periods.

Knowledge Spillovers from Companies with Different Roles in the Industry Chain over Time

Through observation of the network diagrams over three different periods, we can understand the knowledge spillover performance of companies based on their different roles in the industry chain. IDM's were located at central positions in the network during the 6-inch and 8-inch wafer periods, while foundries began to appear in central positions in the 8-inch wafer period. By the time of the 12-inch wafer, the central positions in the network were mostly occupied by IDM's and foundries. As for IC design companies, packaging & testing companies, IC equipment companies and others, there are a few companies that occasionally approach the central positions of the network. Table 5 shows the results of a multiple comparison analysis of network centrality between companies with different roles in the industry chain; only significant post-
hoc test results are presented. Firstly, considering the strong ties and weak ties of knowledge spillovers, the network centrality of IDM s are significantly higher than IC design companies, foundries, IC equipment companies and others. That is to say, IDMs make more connections through both strong strong ties and weak ties of the knowledge spillover network with other companies than the other five roles in the industry chain during the three periods. Secondly, foundries participate in knowledge spillover activities mostly through strong ties in the 6-inch wafer period. Lastly, the knowledge spillover activities of others are significantly higher than packaging & testing companies through weak ties in the 8-inch period.

**Table 4 ANOVA Test of the Average Number of Connections between Types of Companies based on Technology Position**

<table>
<thead>
<tr>
<th></th>
<th>6-inch wafer</th>
<th>8-inch wafer</th>
<th>12-inch wafer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Strong Ties of Knowledge Spillovers</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Links with Technology Leaders (L)</td>
<td>L &gt; I : B &gt; F, I</td>
<td>L &gt; F, I</td>
<td>L &gt; F &gt; I</td>
</tr>
<tr>
<td>Links with Technology Brokers (B)</td>
<td>L &gt; B</td>
<td>L &gt; F, I : B &gt; I</td>
<td>L &gt; B, F, I</td>
</tr>
<tr>
<td>Links with Technology Followers (F)</td>
<td>L &gt; I</td>
<td>L, B &gt; F, I</td>
<td>L &gt; B, F &gt; I</td>
</tr>
<tr>
<td>Links with Technology Isolated Companies (I)</td>
<td>-</td>
<td>-</td>
<td>L &gt; F, I</td>
</tr>
</tbody>
</table>

**Weak Ties of Knowledge Spillovers**

<table>
<thead>
<tr>
<th></th>
<th>6-inch wafer</th>
<th>8-inch wafer</th>
<th>12-inch wafer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow-out from Technology Leaders (L)</td>
<td>L, F &gt; I</td>
<td>L, F &gt; B, I</td>
<td>L &gt; I</td>
</tr>
<tr>
<td>Flow-out from Technology Brokers (B)</td>
<td>L &gt; B</td>
<td>L, B &gt; F, I</td>
<td>L &gt; F &gt; I</td>
</tr>
<tr>
<td>Flow-out from Technology Followers (F)</td>
<td>L, B &gt; F, I</td>
<td>L &gt; B, F &gt; I</td>
<td>L &gt; F, I</td>
</tr>
<tr>
<td>Flow-out from Technology Isolated Companies (I)</td>
<td>L, B &gt; F, I</td>
<td>B &gt; L, F &gt; I</td>
<td>L &gt; F &gt; I</td>
</tr>
</tbody>
</table>

Note: L=Technology Leaders, B=Technology brokers, F=Technology followers, I=Technology Isolated Companies. The mean difference is significant at the .05 level.

**Table 5 ANOVA Test of the Network Centrality of Companies based on the Different Roles in the Industry Chain**

<table>
<thead>
<tr>
<th></th>
<th>6-inch wafer</th>
<th>8-inch wafer</th>
<th>12-inch wafer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Strong Ties</strong></td>
<td>IDM &gt; ICd, Eqpt, Oths Fdry &gt; Eqpt</td>
<td>IDM &gt; ICd, Pakg, Eqpt, Oths</td>
<td>IDM &gt; ICd, Pakg, Eqpt, Oths</td>
</tr>
<tr>
<td></td>
<td>IDM &gt; Fdry, ICd, Oths</td>
<td>IDM &gt; ALL Oths &gt; Pakg</td>
<td>IDM &gt; ALL</td>
</tr>
</tbody>
</table>

Note: IDM=IDM, Fdry=Foundry, ICd=IC Design Companies, Pakg=Packaging & Testing Companies, Eqpt=IC Equipment Companies, Oths=Others. The mean difference is significant at the .05 level.

Table 6 represents the results of a multiple comparison analysis testing the average number of connections between one type of company and another based on their roles in the industry chain; only significant post hoc results are presented. There are a total of 21 possible pairings in the undirected strong ties network and 36 possible pairings in the directed weak ties network. This way, we can examine which two categories of companies more frequently link to each other in the different types of knowledge spillover networks. The results show that through the strong ties of knowledge spillovers, IDMs on average make significantly more connections with other types of companies than those with the other five roles in the industry chain. During the 6-inch wafer period, IDMs and IC design companies built significantly more R&D connections with IDMs than companies with other roles in the industry chain. In the 8-inch wafer period, IDMs, IC design companies and other companies all had significantly more R&D cooperation activities with IDMs than that with other types of companies. Moving to the 12-inch period, all
types of companies excepting IC equipment companies had a significantly higher number of connections with IDMs than with the other roles in the industry chain.

**Table 6 ANOVA Test of the Average Number of Connections between Types of Companies based on Role in the Industry Chain**

<table>
<thead>
<tr>
<th></th>
<th>6-inch Wafer</th>
<th>8-inch Wafer</th>
<th>12-inch Wafer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Strong Ties of Knowledge Spillovers</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Links with IDM</td>
<td>IDM&gt; Fdry, Pakg, Eqpt, Oths</td>
<td>IDM&gt; Icd, Pakg, Eqpt, Oths</td>
<td>IDM&gt; Icd, Pakg, Eqpt, Oths</td>
</tr>
<tr>
<td>Links with Foundry</td>
<td>-</td>
<td>-</td>
<td>IDM&gt; Icd, Pakg, Eqpt, Oths</td>
</tr>
<tr>
<td>Links with IC Design</td>
<td>IDM&gt; Icd, Pakg, Eqpt</td>
<td>IDM&gt; Icd</td>
<td>IDM&gt; Icd</td>
</tr>
<tr>
<td>Links with Packaging &amp; Testing</td>
<td>-</td>
<td>-</td>
<td>IDM&gt; Eqpt, Oths</td>
</tr>
<tr>
<td>Links with IC Equipment</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Links with Others</td>
<td>-</td>
<td>IDM&gt; Pakg</td>
<td>IDM&gt; Icd</td>
</tr>
</tbody>
</table>

| **Weak Ties of Knowledge Spillovers** |              |              |               |
| Flow-out from IDM    | IDM> Icd, Eqpt | IDM> Icd, Eqpt, Oths | IDM> Icd, Eqpt, Oths |
| Flow-out from Foundry| -            | Fdry> IDM, Icd, Pakg, Eqpt, Oths | Fdry> Icd, Pakg, Eqpt, Oths |
| Flow-out from IC Design | - | - | IDM> Pakg, Eqpt, Oths |
| Flow-out from Package | - | - | IDM> Icd, Eqpt, Oths |
| Flow-out from IC Equipment | - | IDM> Pakg, Oths | IDM> Icd, Pakg, Oths |
| Flow-out from Others  | -            | IDM> Icd, Pakg, Eqpt, Oths | IDM> Fdry, Icd, Pakg, Eqpt, Oths |

Note: IDM= IDM, Fdry= Foundry, Icd= IC Design, Pakg= Packaging & Testing, Eqpt= IC Equipment, Oths= Others. The mean difference is significant at the .05 level.

Next, results from the weak ties of knowledge spillovers show that IDMs are on average significantly more likely to be the receivers of knowledge spillovers than the other types of companies. In the 6-inch wafer period, IDMs received significantly more connections from other IDMs than from other roles in the industry chain. During the 8-inch wafer period, a significantly higher number of weak ties flowed from IDMs, IC equipment companies and other companies to IDMs than from other types of companies. Entering the 12-inch wafer period, all types of companies had significantly more knowledge spillovers through weak ties to IDMs than to other roles in the industry chain. Furthermore, the number of weak tie relations was significantly higher from foundry to foundry, from IC design companies to foundries and to other IC design companies, and from packaging & testing companies to other packaging & testing companies.

**Conclusions**

*Technology Leaders are the Main Receivers of Semiconductor Knowledge Spillovers among all Technology Positions*

This study analyzed the differences between types of receivers of knowledge spillovers using four categories of technology positions in the semiconductor industry. We found that technology leaders are the major receivers in both the strong ties and weak ties of knowledge.
spillover networks during all three periods. Particularly, semiconductor knowledge spilled through strong ties between all types of companies and technology leaders, except for that of technology isolated companies during the 6-inch and 8-inch wafer periods. As for the weak ties network of knowledge spillovers, knowledge primarily flowed from other types of companies to both technology leaders and technology brokers during the 6-inch and 8-inch wafer periods. Technology leaders became the major receivers of knowledge through the weak ties of knowledge spillovers by the time of the 12-inch wafer period. However, the knowledge significantly flows out from all types to leaders, it not means leaders’ R&D depends on other types.

*IDMs and Foundries are the Major Receivers of Knowledge Spillovers in the Semiconductor Industry among all Roles in the Industry Chain*

In order to discuss the different types of knowledge spillover receivers in the semiconductor industry, this study divided companies into six categories based on their roles in industry chain. The results show that IDMs are the primary receivers in both the strong ties and weak ties of knowledge spillovers networks over the three wafer diameters periods. In addition, foundries have become major receivers of both strong ties and weak ties of knowledge spillovers during the 8-inch and 12-inch wafer periods. Focusing on the 6-inch wafer period, IDMs were the dominant participants of R&D activities in the semiconductor industry, while IC design companies and foundries accounted for only a very small fraction of knowledge spillover connections. However, foundries have increased in importance due to the specialization and division of labour within the semiconductor industry. By the time of the 12-inch wafer period, several IDMs have outsourced all of their wafer production to foundries and have thus transformed to purely IC design companies. Reflecting such trends, which include the increase in participation of foundries and the role transformation of IDMs, the recipients of semiconductor knowledge spillovers have changed as well. The division of labour within the semiconductor industry is dynamic, and many IDMs have converted to fabless companies; in other words, their roles have gradually transformed to professional IC design in the industry chain. Hence, the relationship between IDMs and other companies with different roles in the industry chain have also transformed from competitive to cooperative. Moreover, we may conclude that the relationship between IDMs and foundries are getting closer based on the increasing number of knowledge spillover connections between each other.

*Overseeing the Strategy of Semiconductor Technology Development from the R&D Dynamics of Technology Leaders*

This paper measures the involvement of companies in knowledge spillover activities and concludes that technology leaders are located in central positions in the network among all different technology positions. IDMs and foundries are major participants in semiconductor technology research and development among companies with various roles in the industry chain. Technology leaders, IDMs and foundries all have high involvement in both strong ties and weak ties of knowledge spillovers, and play the role of knowledge providers in the semiconductor industry. As a result, we suggest that interested parties may monitor the technology developments of an industry by filtering out companies that are both technology leaders and IDMs or both technology leaders and foundries, to be used as references in the formulation of R&D strategies.

*The Relationship between Foundries and IDMs have changed from Vertical Outsourcing to Horizontal Cooperation*

The rise of foundries is due to the need to compensate for the insufficiencies in production capabilities of IDMs. As the semiconductor industry evolved into highly specialized divisions
of labour, many IDM s have transformed into fabless companies and focus on their expertise in IC design. AMD is one representative of such transformation. On the other hand, foundries have become in charge of R&D in wafer production. This study found that beginning in the 8-inch wafer period, foundries have become one of the major receivers of strong ties and weak ties of knowledge spillovers, second to IDMs. In other words, foundries have become major participants in the strong ties and weak ties of knowledge spillovers. We can further observe that in 12-inch wafer period, the patent count of TSMC (a foundry) has reached AMD (an IDM) in our case study. What is more, the involvement of TSMC in the strong ties of knowledge spillovers is even higher than IDMs such as AMD and Samsung. Hence, we conclude that the relationships between foundries and IDMs are no longer vertical contract manufacturing but horizontal R&D cooperation.

Suggestions for Future Research
This research applied two indices—technological crowding and technological prestige to further divide semiconductor companies into technology leaders, technology brokers, technology followers and technology isolated companies. As the development of wafer diameters increases, the technology position of a company might change. In the analysis of the changes in a company’s technological position, we suggest to not only compare a company’s technological position with its role in the industry chain, but to also examine the differences of performances between companies whose technological positions have changed with those whose technological positions remained unchanged. Thus we may more fully investigate how the variation in technological positions may influence the performance of companies in knowledge spillovers activities.

References
The Missing Link Effect: A Case Study Using Patent Main Path Analysis

Chung-Huei Kuan¹ and Dar-Zen Chen²

¹maxkuan@mail.ntust.edu.tw
National Taiwan University of Science and Technology, Taipei (R.O.C., Taiwan)

²dzchen@ntu.edu.tw
National Taiwan University, Taipei (R.O.C., Taiwan)

Abstract
Both direct citation and Bibliographic Coupling (BC) are valid tools in detecting similarity or relatedness among patents. This study is then motivated by the question: what is the difference between a direct citation and a strong BC but without direct citation? We found that the latter, referred to as missing link here, seems to be useful in capturing the technology relatedness among patents filed or issued at close times where direct citations are less likely to be established. We then applied the Main Path Analysis to a patent citation network involving 38,652 patents/publications, 188,721 direct citations, and 9,213 missing links, and derived two main paths with and without the missing links added. By comparing these main paths, we found that the missing links, merely accounting for less than 5% of all connections in this case, may identify a series of intermediate and related patents reflecting elaborative evolvement, which may well be overlooked when only direct citations are considered.

Conference Topic
Patent analysis

Introduction
A first patent is cited or referenced by a second, subsequent patent because the applicant of the second patent considers that the first patent provides, addresses, or discloses similar or related technical background, issue, or solution. The first patent may also be cited by the examiner of the second patent as the first patent may challenge the patentability of the second one, in addition to those reasons by the applicant.

Patent citations have long been accepted as having high bibliometric value (cf. Hall, Jaffe, & Trajtenberg, 2005; Jaffe, Fogarty, & Banks, 1997; Trajtenberg, 1990). In many applications, they are assumed to have revealed the similarity or relatedness between patents.

A representative research area based on the above assumption is the knowledge or technology spillover where knowledge flow between similar or related patents associated with different geographic areas or industries is observed. Here we list a small number of studies in this area published after 2013, demonstrating that utilizing patent citations as a proxy for knowledge flow is a widely accepted approach. Murata et al. (2014) confirmed knowledge spillover are significantly localized using cited-citing relationships between U.S. patents granted between 1975 and 1999. Figueiredo, Guimarães, and Woodward (2015) found that knowledge spillover is positively correlated with industry localization and that the localization of an industry may offset the adverse effect of distance, using a set of cited-citing pairs of U.S. patents. Li (2014) studied the effects of distance and subnational/national borders on international and intranational knowledge spillovers through patent citations across some most cited countries and metropolitan statistical areas within U.S. and claimed that border and distance effects increase over time, but fade as patents age. Karvonen and Kässi (2013) collected patent data from main players of the radio-frequency identification (RFID) value chain in EPO database,
and used citation data to investigate the emergence of new industry segment. Kim, Lee, and Sohn (2016) quantified the spillover of unmanned aerial vehicle (UAV) technology to various industries by retrieving U.S. patents that cited UAV patents and matching these citing patents to industries. Noailly and Shestalova (2017) used citations of European patents to see what technologies are built on the knowledge developed in renewable energy.

Researchers have been comparing applicant-submitted references (hereinafter, applicant citations) and examiner-located references (hereinafter, examiner citations) for their difference. Criscuolo and Verspagen (2008) found that examiners more often provide citations that may compromise patentability, and that applicants tend to ignore prior art that may endanger their patent applications. Hegde and Sampat (2009) found that examiner citations are better value indicator as they have a much stronger relationship with patent renewal probability than the applicant citations do. Cotropia, Lemley, and Sampat (2013) argued that patent examiners rely almost exclusively on prior art they find themselves, instead of applicant citations, in rejecting patent applications or in limiting patent scopes. In contrast, Thompson (2006) indicated that applicant citations are more relevant as applicants are more familiar with their inventions than the examiners are. Alcacer, Gittelman, and Sampat (2009) suggested that applicant citations express knowledge flows better than examiner citations. Park, Jeong, and Yoon (2017) found that the patents cited by applicants have higher quality than those cited by examiners in four industries.

A patent citation, whether it is from the applicant or the examiner, indeed reflect a connection between the cited and the citing. The difference seems to lie only in the degree of similarity, relatedness, or how much knowledge indeed flows from the cited to the citing. What intrigues us here is not the origin of the citation, but the absence of direct citation between two seemingly related patents.

The “missing” citation

In addition to the utilization of direct citations outlined above, another approach in detecting or measuring the similarity or relatedness between two documents or patents is Bibliographic Coupling (BC). Two documents or patents are bibliographically coupled if they both cite at least one common reference. The Bibliographic Coupling Strength (BCS) represents the number of common references. Kessler (1963) proposed the concept Bibliographic Coupling and used it to measure subject similarity. The concept then quickly drew significant interest from researchers. Various applications have been designed and we listed a small number of publications in using BC to detect or measure the similarity or relatedness of patents as follows. Wartburg, Teichert, and Rost (2005) proposed a multi-stage patent citation analysis for the measurement of inventive progress using BC. Park et al. (2015) used BC to measure technological similarity in the fuel cell membrane electrode assembly technology so as to locate potential R&D collaboration partners. Huang and Chang (2014, 2015) applied BC with sliding window to track the generation, growth, decline, and disappearance of research fronts.

If both direct citation and BC reflect the similarity or relatedness between two patents, it is interesting to notice that sometimes there is no direct citation between two patents that are strongly bibliographically coupled.

Taking two U.S. utility patents, US8,622,222 and US8,623,202 as an example, they are both related to membrane bioreactor technologies, were both filed by the same company, one in January 2011 and the other October 2012, but with different inventors. They were both granted
within January 2014 (therefore, their patent numbers are very close) after being examined by different examiners. The two patents do not have direction citation between them, but their similarity is clearly reflected by their exceptionally high BCS 1,039 (US8,622,222 has total 1,109 references and US8,623,202 has 1,108 references).

Taking another two more distant U.S. utility patents, US8,585,882 and US9,586,842, as an example, they are both related to water treatment, but filed by different companies, one in December 2008 and the other December 2015. They are granted in November 2013 and March 2017, respectively, by different examiners. The two patents do not have direction citation, but their similarity is also clearly reflected by their high BCS 465 (US8,585,882 has 558 references and US9,586,842 has 622 references).

The term missing link is often used loosely to describe a newly found evidence indicating a possible evolution path between a predecessor and a successor such as a new piece of fossil manifests a possible ancestor to the human lineage. We borrowed the term here to denote this latent connection between patents suggested by their strong BC.

Some researchers had already noticed the occurrence of missing links and made use of them. Chen et al. (2011) used missing links to construct more comprehensive technological clusters using light emitting diode (LED) as a case study. Yeh et al. (2013) took one step further by filtering out direct citations between patents with weak BC, in addition to incorporating missing links into a patent citation network.

Why do missing links occur? One possible explanation for the lack of direct citation is that applicant citations or examiner citations are by no means exhaustive, and it is surely possible that the applicant or examiner of one patent fails to learn or discover the presence of the other.

On the other hand, Chen et al. (2011) calculated the time lags for missing links and found that, for a majority of them (65.52%) in their case study, one has application date before the other’s issued date (i.e., their application processes are overlapped in time). The authors then argued that the later patents’ applicants at the times of filing could not be aware of and cite the earlier ones since information about the earlier ones was not available to the public then.

We are then motivated by the question: what is the difference between a direct citation and a strong BC but without direction citation? Prompted by Chen et al. (2011), the answer seems to lie in the timing characteristics of missing links, and an empirical study was conducted for this purpose.

**Empirical Data**

We chose the technical area of carbon dioxide (CO₂) capture, storage, recovery, delivery, and regeneration as our empirical case. We collected the related utility patents and their references (including U.S. patents and publications) all issued or published after 1976 and before 2017/03/28 from USPTO database. These patents/publications have at least one of some specific keywords in at least one of the relevant fields (i.e., title, abstract, specification, and claims) and also have at least one of some specific Cooperative Patent Classification (CPC) symbol prefixes. There are then total 38,652 patents and publications with 188,721 pairs having direct citations and 1,609,549 bibliographically coupled pairs.
The 1,609,549 BC pairs have a significantly skewed BCS distribution with a mean 2.74 (μ), a standard deviation 15.66 (σ), and a maximum 1,123. Among them, 1,477,783 pairs, accounting for 92% of all BC pairs, have BCS not greater than 3 (the closest integer to the mean BCS). Since missing links do not actually exist, we decided to be conservative and considered only BC pairs having BCS greater than 34 (=μ+2σ). There are only 13,013 such pairs, accounting for only 0.81% of all BC pairs. We then further removed 3,800 of them that already have direct citations. Finally we were left with as few as 9,213 (=13,013-3800) BC pairs having a range of the BCS between 35 and 1,123. The BCS distribution of these BC pairs is shown in Figure 1. A great majority of the BC pairs (8,674 or 94%) have BCS less than or equal to 300.

![Figure 1. The BCS distribution of the 9,213 BC pairs (vertical axis is in log scale).](image)

Figure 2 shows the distributions of time spans between the filing dates (in lighter bars) and between the issued dates (in darker bars) for the 9,213 BC pairs. As illustrated, 58% of the 9,213 BC pairs have filing date differences fall within 2 years, and 65% have issued date differences fall within 2 years. Then, as the time span between patents’ filing and issued dates increases, it seems that they are less likely to form strong BC pairs as manifested by the BC pairs’ decreasing trends to the right shown in Figure 2.

This seems to confirm the observation of Chen et al. (2011) that, if two patents are filed or issued within a short time span, their applicants or examiners are more difficult to be aware of each other, as both patents may remain undisclosed before filing or remain pending in the patent office before issuance.

Overall, the above observation seems to suggest that the missing links would be most useful in capturing those technology relatedness among patents where they are filed or issued at close times and where direct citations are less likely, in not impossible, to be established.

To further investigate whether our speculation is reasonable, we conducted the so-called Main Path Analysis for our empirical case.
Main Path Analysis

Main Path Analysis (MPA) was first taught by Hummon and Doreian (1989) to discover the major development trajectory of a scientific field through identifying the most significant chain of citations in a citation network. This method generally involves the following steps: (a) constructing a citation network from relevant documents of the scientific field; (b) calculating a weight for each link of the citation network related to the link’s traversal counts; and (c) searching for a series of connected links across the network according to their weight. This series of links is then referred to as the main path of the scientific field, and considered as having embodied a development trajectory of the field.

Since its inception, researchers have been applying this method in, for example, detecting technological changes, knowledge transformation (Lucio-Arias & Leydesdorff, 2008; Martinelli, 2012; Mina et al., 2007), and reviewing literatures (Bhupatiraju et al., 2012; Calero-Medina & Noyons, 2008; Colicchia & Strozzi, 2012; Harris et al., 2011; Liu et al., 2013; Lu, Hsieh, & Liu, 2016). The same method is also applied to patent citation networks in mapping technological evolvements (Fontana, Nuvolari, & Verspagen, 2009; Park & Magee, 2017; Verspagen, 2007). MPA capability is built in the well-known network analysis application Pajek (Batagelj & Mrvar, 1998).

The rationale of MPA is to consider each link of the citation network as a path of knowledge flow from the cited to the citing. Then, an algorithm is applied to determine a weight for each link related to the number of times the link is traversed. For example, Figure 3 is a fictitious citation network where the link weights are determined using the algorithm Search Path Link Count (SPLC) (Hummon & Doreian, 1989). The weight of the link 5→6 (i.e., from node 5 to node 6) is 15, because it counts the number of traversals of the link 5→6 from all preceding nodes (1 to 5) to the sink nodes (7, 8, 9). Similarly, the weight of the link 6→7 is 6 because there are 6 preceding nodes (1 to 6) and each traverses the link 6→7 once to reach the sink node 7.
There are other algorithms for weight calculation such as SPC (search path count) (Batagelj & Mrvar, 1998), SPNP (Search Path Node Pair) (Hummon & Doreian, 1989), etc. These algorithms all determine the weight solely based on the topological location of the link, and a link would have a greater weight if it can be reached from more nodes or more nodes may be reached through it. The link weight based on the traversal count therefore functions as a proxy to the total amount of knowledge flow through the link.

There are also different methods in determining the main path once the link weights are set. The main path of Figure 3 is denoted by black nodes using global search method (Liu and Lu, 2012) which selects a path with the greatest total weight from source to sink nodes. The local search method starts from the source or sink nodes and selects the link(s) from/to the source/sink nodes with the greatest weight(s) and works forward/backward for the next search until a sink/source node is reached (Hummon & Doreian, 1989). The key-route method (Liu and Lu, 2012) determines one or more main paths by locating the link(s) with the greatest weight first and tracing backward and forward until a source or sink node is reached.

By treating missing links as real citations and applying MPA, we are able to see whether the missing links would lead to any difference in the resulted main path. We then may gain some insight into the effect of missing links if some major difference does occur.

Main Paths
A patent citation network (PCN) was constructed out of our empirical case and involved 38,652 nodes and 188,721 arcs. We then adjusted the original patent citation network (OPCN) by incorporating 9,213 missing links as real citations. Each missing link always originates from the lower numbered patent to the higher numbered patent. As there are 188,721 links in the OPCN, these missing links account for only 4.6% (=9,213/ (188,721+9,213)) of the links of the adjusted citation network (APCN).

In determining the main path, we chose to use the SPLC algorithm in calculating the link weights as prior researches had reported that algorithms SPC, SPLC, SPNP, etc. all derive similar results (cf. Batagelj, 2003; Verspagen, 2007). Further, we chose to use the global search method (Liu & Lu, 2012) in developing the main path. The key-route method is a recent variant to the MPA, but it will usually produce more than one main path. To simplify the comparison, we therefore chose the global search method as it will usually produce a single main path having the globally greatest total weight.
The original main path derived without missing links and the adjusted main path obtained with the missing links supplemented are displayed together in Figure 4 using nodes and arcs of different styles for easier comparison. There are total 67 different nodes (patents) from both the original and adjusted main paths, and they are numbered from 1 to 67 in ascending order of their patent numbers and therefore their issued dates. A complete list of these 67 patents is provided in the Appendix.

There are 19 black nodes representing those patents present on both the original and adjusted main paths. There are 8 white nodes representing those appearing only on the original main path, and the 41 grey nodes are those present only on the adjusted main path. The solid grey arcs are direct citations, and the dashed black arcs are missing links.

The original main path involves 27 patents and is represented by the chain of black and white nodes connected by solid grey arcs (direct citations). The adjusted main path contains total 59 patents and is represented by the chain of black and grey nodes connected either by solid grey arcs (direct citations) or dashed black arcs (missing links).

As can be observed from Figure 4, the adjusted main path has captured two third (18) of the patents from the original main path. It seems fair to claim that both the original and adjusted main paths derive a substantially identical trajectory of knowledge flow in our case study, but the adjusted main path has offered significantly more information, as it provides additional 41 (=59-18) patents and additional 49 links in the adjusted main path. For the 49 newly added links, 6 of them are direct citations, whereas the other 43 are all supplemented missing links.

Based on where the newly included nodes and arcs occur, the adjusted main path may be partitioned into 4 sections, and their end nodes and span years (based on end patents’ issued years) are listed in Table 1. These 4 sections allow us to examine more closely the effect of missing links.

In Section 1, the original and adjusted main paths are completely identical. This is because a patent in an early stage of a technology field has fewer references. For example, the end nodes 1 (US3,977,845 issued in 1976) and 8 (US4,813,980 issued in 1989) have only 7 and 15 references. As we considered only BC pairs having BCS greater than 34, no BC pair within this stage would satisfy the threshold.

In Section 2, the end node 19 (US6,221,117 issued in 2001) has 140 references, indicating that there would be BC pairs satisfying the threshold and missing links are supplemented. These missing links help to identify and add new nodes 10, 11, and 18 to the adjusted main path, but these 3 nodes are all connected by direct citations, not missing links. The reason is that the supplemented missing links are few and dispersed, and they are not able to cooperatively thrust any one of them into a status as a major link of knowledge flow. However, they are able to strengthen some existing direct citations (i.e., 8→10→11→14, and 16→18→19) to become part of the adjusted main path.
In Section 3, the end node 60 (US8,057,575 issued in 2011) has 276 references, implying that there are more missing links incorporated into the APCN in this stage than in the previous stage. In contrast to what is observed from Section 2, these more numerous missing links are able to cascade together into a number of different path segments. We will talk more about Section 3 later.

Section 4 provides a different scenario as the original and adjusted main paths develop separately. We speculate that the development of the original and adjusted main paths is still in progress. It is possible that the original and adjusted main paths may merge at a future patent, just like what happens, for example, at node 25 that merges the original path segment 19→20→23→25 and the new path segment 19→21→22→24→25.

Table 1 also provides the minimum, maximum, and average BCS for the missing links in Sections 3 and 4, respectively. We can see that their BCS is not particularly strong, and actually none of the strongest 5% shown in Figure 1 has emerged as part of the adjusted main path. This
observation does not come as a surprise since, for a missing link to appear on the adjusted main path, it is determined by its topological location, not by its BCS.

Based on the above observation, we can see that there are some key patents denoted by nodes such as 19, 25, 29, 36, 38, 42, 52, and 60 of Section 3 providing a backbone of the knowledge flow as they are present on both the original and adjusted main paths. However, the adjusted main path suggests totally different path segments between these key patents involving completely different sets of patents connected almost entirely by the missing links. This seems to suggest that missing links may indeed deliver more topological significant relatedness than what is captured by the replaced direct citations.

To see that this phenomenon is not coincident, we conducted a number of experiments by manually multiplying the weights of all missing links by 0.5, 0.1, and 0.01 after all link weights are determined and before applying the global search method to determine the main path. In other words, we artificially and purposely reduced the weights of all missing links to one half, one tenth, and one hundredth of their original values. We found that the resulted main paths are exactly the same as the original one. This implies that the weights (i.e., traversal counts) of the 43 missing links on the adjusted main path are so strong that downplaying their importance to a great degree still wouldn’t produce any difference. Therefore, the 43 missing links indeed capture some topologically significant connections among patents.

In addition, please pay special attentions to nodes 25, 29, 36, 38, 42, 52, and 60 in Section 3. These patents are issued at close years 2003, 2003, 2004, 2005, 2006, 2010, and 2011, respectively (except the patents denoted by node 42 and 52 where they are issued at years 2006 and 2010). In the OPCN, they are sequentially cascaded by direction citations but, in the APCN, each of these direct citations is replaced by a series of missing links connecting a number of patents not on the original main path.

Taking nodes 52 and 60 as example, the arc in the original main path reflects the knowledge flow between the patents denoted by the two nodes. But with the help of the missing links, patents denoted by nodes 53 to 59 are manifested in the adjusted main path. As we already know, these are related patents occurring at close times (i.e., all issued between 2010 and 2011) and their relatedness may be overlooked by direct citations. Therefore, it seems that this series of intermediate and related patents may give us a more detailed picture regarding, for the original two patents denoted by nodes 52 and 60, how one evolves to the other.

As mentioned earlier, the main idea behind MPA is to use citations as proxies for knowledge flow. But missing links are not real citations, they are only simulations of real citations in the APCN. Therefore the adjusted main path derived from the APCN should not be considered as a better or more accurate path. The adjusted main path in this study is mainly used as a tool for observing the effect of missing links.

**Discussion and Conclusion**

Both direct citations and BC are valid tools in detecting similarity or relatedness, or in capturing knowledge flow among patents. Then, we were intrigued by a conflicting scenario that two patents lack direct citation but are strongly bibliographic-coupled.

Based on empirical data, we found that a significant portion of missing links have their patents’ filing or issued dates fall within one to two years. In other words, these patents are prosecuted
concurrently, and it would be difficult, if not impossible, for their applicants or examiners to become aware of and cite each other. Yet their relations may well be captured by missing links. Therefore, missing links seem to be useful in capturing the technology similarity or relatedness among these concurrent patents while direct citations among them are less likely to be established.

This study then applied MPA to a patent citation network involving 38,652 patents/publications, 188,721 direct citations, and 9,213 missing links. We then derived an original main path without the missing links added, and an adjusted main path with the missing links added to simulate real citations.

We found that some direct citations in the OPCN are replaced by a series of newly found patents connected by missing links in the adjusted main path. We verified that this is not coincident by manually downplaying the weights of the missing links to as low as one hundredth of their original values, and the adjusted main path is not affected at all.

On one hand, this observation suggests that missing links may indeed deliver, from the PCN and main path’s point of view, topologically more significant connections between patents.

More importantly, as a single direct citation of the OPCN is elaborated into a chain of newly found patents in the APCN, we found that these newly found patents not only may fill the gap left by direct citations as described above, but also may give us a more detailed picture regarding, for the two patents connected by the direct citation, how one evolves to the other.

This study may be challenged that an absolute BC threshold (34) is adopted in determining the missing links, and one may argue that a normalized or relative BC threshold should be more appropriate. By using an absolute threshold, indeed we may filter out some important BC pairs having fewer references (like what we observed in Section 1 of Figure 4), and may include some not-so-important BC pairs having more references (e.g., those added in Sections 3 and 4 of Figure 4). It is definitely interesting to see if the normalized or relative BC threshold would contribute any difference. We however were not so concerned in this study since MPA itself has an inherent tendency to filter out those irrelevant links.

This study may be extended along several directions. One extension is that, instead of adding the missing links, one can use BC to filter out irrelevant direct citations and see what difference may occur in the resulted main path. Another extension is that, similar to what Persson (2010) suggested, instead of using BC to determine latent connections, the similar and popular approach Co-citation (CC) may be applied to compare the two approaches.
References


<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3977845</td>
<td>Y</td>
<td>Y</td>
<td>36</td>
<td>6824593</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>4021210</td>
<td>Y</td>
<td>Y</td>
<td>37</td>
<td>6869707</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4077779</td>
<td>Y</td>
<td>Y</td>
<td>38</td>
<td>6953497</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>4</td>
<td>4171207</td>
<td>Y</td>
<td>Y</td>
<td>39</td>
<td>6967063</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4512780</td>
<td>Y</td>
<td>Y</td>
<td>40</td>
<td>6994927</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>4624841</td>
<td>Y</td>
<td>Y</td>
<td>41</td>
<td>7005113</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4671893</td>
<td>Y</td>
<td>Y</td>
<td>42</td>
<td>7052530</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>8</td>
<td>4813980</td>
<td>Y</td>
<td>Y</td>
<td>43</td>
<td>7135048</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>4913709</td>
<td>Y</td>
<td></td>
<td>44</td>
<td>7195663</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>4915711</td>
<td>Y</td>
<td></td>
<td>45</td>
<td>7368194</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>5073356</td>
<td>Y</td>
<td></td>
<td>46</td>
<td>7410531</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>5133785</td>
<td>Y</td>
<td></td>
<td>47</td>
<td>7470293</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>5332424</td>
<td>Y</td>
<td></td>
<td>48</td>
<td>7601302</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>5435836</td>
<td>Y</td>
<td></td>
<td>49</td>
<td>7632322</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>5562754</td>
<td>Y</td>
<td></td>
<td>50</td>
<td>7682718</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>5705916</td>
<td>Y</td>
<td></td>
<td>51</td>
<td>7736596</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>5861137</td>
<td>Y</td>
<td></td>
<td>52</td>
<td>7789941</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>5997594</td>
<td>Y</td>
<td></td>
<td>53</td>
<td>7819955</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>6221117</td>
<td>Y</td>
<td></td>
<td>54</td>
<td>7828864</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>6319306</td>
<td>Y</td>
<td></td>
<td>55</td>
<td>7939051</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>6375906</td>
<td>Y</td>
<td></td>
<td>56</td>
<td>7972420</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>6376113</td>
<td>Y</td>
<td></td>
<td>57</td>
<td>7981172</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>6458189</td>
<td>Y</td>
<td></td>
<td>58</td>
<td>8021446</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>6494937</td>
<td>Y</td>
<td></td>
<td>59</td>
<td>8038748</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>6537352</td>
<td>Y</td>
<td></td>
<td>60</td>
<td>8057575</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>6562111</td>
<td>Y</td>
<td></td>
<td>61</td>
<td>8157900</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>6569227</td>
<td>Y</td>
<td></td>
<td>62</td>
<td>8257466</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>6596057</td>
<td>Y</td>
<td></td>
<td>63</td>
<td>8636828</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>6632270</td>
<td>Y</td>
<td></td>
<td>64</td>
<td>8691463</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>6641625</td>
<td>Y</td>
<td></td>
<td>65</td>
<td>8696772</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>6719831</td>
<td>Y</td>
<td></td>
<td>66</td>
<td>8961627</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>6719832</td>
<td>Y</td>
<td></td>
<td>67</td>
<td>9187324</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>6723156</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>6767389</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>6783741</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Co-citation analysis based on the position: a case study of Journal of Informetrics

Rongying Zhao¹, Wei Quan², Fengjiao Guo³, Zhuheng Fu⁴
¹ zhaorongying@126.com
Research Center for Chinese Science Evaluation of Wuhan University,
School of Information Management, Wuhan University, Wuhan, China
² xiaoweiweijulia@126.com
Research Center for Chinese Science Evaluation of Wuhan University,
School of Information Management, Wuhan University, Wuhan, China
³ happyguofj@126.com
Research Center for Chinese Science Evaluation of Wuhan University,
School of Information Management, Wuhan University, Wuhan, China
⁴ 510753750@qq.com
School of Foreign Languages,
Wuhan University of Technology, Wuhan, China

Abstract
According to traditional co-citation analysis, co-citation frequency is mainly based on bibliographic references to calculate, which does not consider the multiple co-citation and the relative position of co-citation in a same academic publication. This study aims to thorough mining the essential characteristics of co-citation, in fact, co-citation position and distance, by constructing a position-based co-citation analysis framework and performing an empirical study. The results suggest that the distribution of citations in the text are extremely uneven, most of the citations are concentrated at the front of the article, as well as clustered, and also citation distribution is according to document types; the similarity of academic publications is not only concerned with citation frequency, but also concerned with co-citation level, the result of traditional co-citation analysis covers the effect of non-paragraph level; the co-citation probability in different level is decreasing with the increasing of co-citation frequency, co-citation frequency and the co-citation level have a negative linear relationship; the position-based co-citation analysis can enhance the co-citation clustering and benefit to deep research and evaluation.

1. Introduction
Co-citation analysis was respectively proposed by Small and Marshakov in 1973(Qiu and Ma, 2008), because of the objectivity of method, the validity of the data and the scientific principles of classification (Liu, 2005), co-citation analysis attracts the attention of the Scientometric community (Liu and Zhang, 2013). However, according to the traditional co-citation analysis, co-citation frequency is mainly based on bibliographic references to calculate (Leydesdorff, 2008). It does not consider the multiple co-citation and the relative position of co-citation in a same academic publication. Full text data of scientific literature contained the specific citations, such as the number of citations, citation position, and author's citation motivation. Therefore, co-citation analysis which based on the full text can explore the essence of co-citation. In recent years,
the continuous development and improvement of information structure technology, natural language processing technology and semi-structural information organization technology, bring new opportunities to rise and development of the full text co-citation analysis.

2. Related work
A review of the literature reveals that co-citation analysis has gradually increasing, including correlatively theoretical research (Small, 1980; Gmür, 2003), research progress (Wang and Leng, 2006), system design and development (Cui and Hu et al., 2000), comparative studies with other methods (Boyack and Klavans, 2010) and application research (White and Mccain, 1998). However, full text co-citation analysis research is very limited, and research on the relationship between the co-citation distance and the similarity of co-citation literature is still under theoretical level, lacking of empirical research and cases study.

Elkiss(2008) used quantitative analysis found that citation summaries overlap to some extent with the abstracts of the papers and that they also differ from them in that they focus on different aspects of these papers than do the abstracts (Elkiss and Shen et al., 2008). Gipp and Joran (2009) under the assumption that closer citations are to each other in a document, the more likely it is that the cited documents are related, then they divided co-citation into 5 levels, citations listed in the(1) same sentence, (2) same paragraph, (3) same chapter, (4) same journal, (5) in different versions of the same journal, and define CPI (Citation Proximity Index) as 1, 1/2, 1/4, 1/8, 1/16 (Gipp and Beel, 2009). Liu and Chen (2012) use the proximity of co-citation to describe the co-citation documents’ references relative position distance, from four levels of sentences, paragraphs, chapters, articles, discussing the co-citation distance of the full text scientific publications, and statistics of 22 kinds of Open Access Journals of the co-citation frequency and co-citation position (Liu and Chen, 2012). Liu et al. (2015) proposed an improved algorithm based on citation context, co-citation document and document coupling (Liu and Yu et al., 2014). Boyack and his colleagues (2013) proposed to improve the accuracy of co-citation clustering through full-text data, they define co-citation distance as two co-citation position between the number of bytes the number of bytes representing percentage, divide co-citation hierarchy as (1) in the same position, (2) less than 5%, (3) 5% to 15%, (4) 15% to 25% and (5) above 25%, respectively, weight is set to 4, 3, 2, 1 and 0. Hu Zhigang, who distinguishes differences and connections between full-text based co-citation analysis based on the bibliographic data (Hu and Chen, 2013). Liu Shengbo, according to the position of two references cited in a document, co-citation relations can divide into sentence level, paragraph level, section level, and article level (Liu and Zhang, 2013).

Research object of position-based co-citation analysis is based on the full-text data of citing articles (Zhao and Zeng, 2014). Classifying and quantifying the individual co-citation strength on co-citation level from the citation position aspect, furthermore, it can reveal and measure the relevance of the co-citation relationship between documents. This study intends to build a position-based co-citation analysis method framework on the basis of previous studies, then through empirical analysis, and mining the relationship between the co-citation distance and co-citation document similarity, and comparing the validity difference between position-based co-citation analysis and traditional co-citation analysis.

3. Methodology
Before the construction of the method framework, we define the co-citation distance and co-citation intensity, co-citation distance refers to the relative distance of the co-citation document in...
the citing document; Co-citation strength is defined as the citation numbers of two articles which cited by other document, in fact frequency, in order to measure the relevance of co-citation documents.

Comparing to the traditional co-citation analysis, the contents and methods of analysis are different in position-based co-citation analysis, first we need to develop co-citation criteria level, the co-citation is divided into different levels respectively, and then analysis the distribution of co-citation position and co-citation level. Secondly, we need to assign the co-citation strength to different levels, and construct the co-citation matrix for the cluster analysis. Position-based co-citation analysis framework, as shown in Figure 1.

![Figure 1 Framework of Position-based Co-citation Analysis](image)

### 3.1 The division of co-citation levels


In a chapter, which has a chapter numbers, chapter title and other information; among the citation information, citation ID and citation context information; in the references section, document ID, title, authors, journals and other information.

Our study intends to use the method which proposed by Liu and Chen etc. (Liu and Chen, 2012; Liu and Zhang, 2013) and from the point of view of the structure of the article, this paper divides the co-citation level into four parts: the sentence level, the paragraph level, the chapter level and the article level.

### 3.2 Assignment of common cited distance

In the traditional co-citation analysis, as long as the two papers were cited at the same time (without considering the times of citation and position), their co-citation strength is 1. The analysis of the existing research illustrate that two articles were cited the closer their co-citation strength (relevance) is bigger, combining with existing research, and in order to facilitate of calculation, the co-citation distance of sentence, paragraph, section are assigned to 4 3, 2, 1. In addition, due to the normative documents, the same paragraphs or sections of the article are generally not the same sentence to express the same meaning of repetition, therefore in the same document in two or more than two times in the same sentence of the cited literature shows that they are in at least two aspects has correlation, the author that its relevance is higher than that of only a sentence level co-citation, so this is a separate named sentence layer two times cited and assigned to 8 (two sentences layers). For the shortest distance in the literature, which is referenced by many different levels in the same document, the shortest distance is taken as an assignment.

According to the above co-citation distance assignment method, select the article *The*
influence of missing publications the on the Hirsch index (Rousseau, 2007) as an example, table 1 is based on the XML data to extract the literature references in the text position, figure 2 is based on the data in Table 1 simplified reference position distribution map, figure 3 is the use of traditional methods and based on the position of the co-citation analysis method to get the co-citation matrix.

<table>
<thead>
<tr>
<th>Reference ID</th>
<th>Sentence</th>
<th>Paragraph</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>sen10</td>
<td>para1</td>
<td>sec1</td>
</tr>
<tr>
<td>A8</td>
<td>sen10</td>
<td>para1</td>
<td>sec1</td>
</tr>
<tr>
<td>A9</td>
<td>sen10</td>
<td>para1</td>
<td>sec1</td>
</tr>
<tr>
<td>A2</td>
<td>sen22</td>
<td>para1</td>
<td>sec1</td>
</tr>
<tr>
<td>A7</td>
<td>sen25</td>
<td>para1</td>
<td>sec1</td>
</tr>
<tr>
<td>A1</td>
<td>sen43</td>
<td>para2</td>
<td>sec1</td>
</tr>
<tr>
<td>A9</td>
<td>sen51</td>
<td>para2</td>
<td>sec1</td>
</tr>
<tr>
<td>A4</td>
<td>sen113</td>
<td>para4</td>
<td>sec2</td>
</tr>
<tr>
<td>A5</td>
<td>sen113</td>
<td>para4</td>
<td>sec2</td>
</tr>
<tr>
<td>A6</td>
<td>sen113</td>
<td>para4</td>
<td>sec2</td>
</tr>
<tr>
<td>A3</td>
<td>sen203</td>
<td>para8</td>
<td>sec3</td>
</tr>
<tr>
<td>A7</td>
<td>sen214</td>
<td>para8</td>
<td>sec4</td>
</tr>
<tr>
<td>A4</td>
<td>sen244</td>
<td>para9</td>
<td>sec4</td>
</tr>
<tr>
<td>A5</td>
<td>sen244</td>
<td>para9</td>
<td>sec4</td>
</tr>
<tr>
<td>A6</td>
<td>sen253</td>
<td>para9</td>
<td>sec4</td>
</tr>
</tbody>
</table>

Figure 2. Single Citation Distribution Chart

Figure 3. Comparison of Traditional Methods (left) and Position-based (right) Matrix

3.3 Data sources and processing

The data of this paper comes from the full-text database of Elsever ConSyn, which provides the full text data of the structured XML format (Hu and Chen, 2012; Hu and Chao, 2013). Selecting the Journal of Informetrics (JOI) as a case study, the journal mainly publish information measurement of the related literature, literature metrology, involving the subject scope is relatively narrow, the relationship between the reference and citation network has a stronger, more conducive to analysis. In the database, the retrieval of all papers published in this journal, through
the "export" function of the database will retrieve the full text of the XML format data, then
download and import EXCEL, extract citation information.

We plan to analysis position-based co-citation from the following two aspects, firstly
statistical analysis of the distribution of co-citation in scientific literature in term of sentence,
paragraph, section, and article; on the other hand, using the above co-citation distance assignment
a citation for single assignment, then two kinds of co-citation matrix for clustering analysis,
comparison of the traditional co-citation and co-citation validity position difference based on the
way.

4. Results and discussions
4.1Position-based co-citation distribution analysis

4.1.1 Reference position distribution

Generally speaking, the scientific paper mainly includes introduction, method, analysis results,
conclusion and discussion. Generally, JOI paper contains 4-6 chapters, including the number of
four chapters, five chapters and six chapters of three papers about the total number of papers 3/4,
we will mainly base on the structure of these three kinds of scientific literature data analysis.

The distribution of citations in each chapter of scientific literature is not uniform, and the
distribution of the number of citations in each chapter is shown in table 2. In the four chapter
structure of the thesis, the first section contains 1285 references, higher than second, section three
or four of the reference number 776, 796 and 271, in the five and six chapters in the paper also
has the same structure characteristics. It can be seen that the reference is mostly distributed in the
front part of the scientific literature, mainly in the first and second sections, accounting for 61.36%
of the total citations.

<table>
<thead>
<tr>
<th>Section 1</th>
<th>Section 2</th>
<th>Section 3</th>
<th>Section 4</th>
<th>Section 3</th>
<th>Section 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section 4</td>
<td>1285</td>
<td>776</td>
<td>796</td>
<td>271</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Section 5</td>
<td>1229</td>
<td>1023</td>
<td>623</td>
<td>540</td>
<td>210</td>
<td>/</td>
</tr>
<tr>
<td>Section 6</td>
<td>756</td>
<td>531</td>
<td>422</td>
<td>234</td>
<td>334</td>
<td>97</td>
</tr>
<tr>
<td>Total</td>
<td>3270</td>
<td>2330</td>
<td>1841</td>
<td>1045</td>
<td>544</td>
<td>97</td>
</tr>
</tbody>
</table>

By analyzing the data of XML full text of each article, we found that there are the following
characteristics in the distribution of references: 1. References are mainly concentrated in the first
section (usually the introduction part). Foreign literature at the beginning of the study need to
have an overview of the background of the study, the previous studies have been summarized and
analyzed, so in the first section of the literature will be relatively more.

Almost all literatures are cited in the first chapter, more than 90% of the references are cited
in the second and third chapters, and in the four or five and six chapters are relatively small. The
two or three chapter will introduce the research methods, the 456 chapter is usually results,
conclusions and discussion part, this part explains in the introduction and reference method is
necessary, scientific research is established on the basis of previous studies, the theory and method
of using the former must be reflected in the article. And in the conclusion and discussion part is
more should reflect their own research results and ideas, the reference is not necessary. （3）The
quote is no single, scholars clustering references in reference, this type of literature is usually the
topic or group by the methods of the same. The clustering cited more persuasive than single quotes,
quoted from the side can explain the different point of view. The different types of distribution of
different reference literature. Review of the citation distribution is relatively uniform, wide range, full text various chapters are basic references, and the method in the methods section of Citation Distribution of the most intensive, ordinary paper in the first part of "Introduction" refers to the most densely distributed. Therefore, we can classify the scientific literature according to the distribution of references, and then study the different types of documents.

4.1.2 Co-citation level distribution analysis

To analyze the references of co-citation distributions, you need to establish a reference of the co-citation relations foundation, marking each co-citation of co-citation level, this process is very complicated, so the data selected from 32 papers published in JOI journals in 2007 as the analysis object. At first, all the references of the 32 literatures were extracted and the weight of these 1138 references were obtained, and 78 references were cited. The cited frequency was greater than or equal to 2. The statistics were 78 references in each document in the co-citation and co-citation level, get the total cited 1699 times (the same as the literature on multiple co-citation count). In which the article layer 569 times, the section level 571 times, paragraph 515 times, the sentence layer 33 times, two times and above the sentence level 11 times. They are accounted for in table 3.

As can be seen from the data in Table 3, the sentence, paragraph and section of the total cited 66.51% of the total citation. Because the data set is based on more than 2 times cited references, the results and analysis of all references based on the results there are some errors in this ratio will be reduced, but it is certain that the sentence level, paragraphs and chapters layer co-citation amount is very large, visible division co-citation of different levels the level of co-citation analysis is necessary to distinguish. The traditional co-citation analysis will be considered as the level of the article, and the results of the analysis cover up the effect of the non-article level.

Table 3. JOI periodicals published in 2007 were cited in the literature of the level of distribution table

<table>
<thead>
<tr>
<th>Traditional</th>
<th>Position-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference number</td>
<td>78</td>
</tr>
<tr>
<td>Co-citation number</td>
<td>1699</td>
</tr>
<tr>
<td>W=1</td>
<td>100%</td>
</tr>
<tr>
<td>W=2</td>
<td>——</td>
</tr>
<tr>
<td>W=3</td>
<td>——</td>
</tr>
<tr>
<td>W=4</td>
<td>——</td>
</tr>
<tr>
<td>W=8</td>
<td>——</td>
</tr>
</tbody>
</table>

4.2 The relationship between the frequency of co-citation and co-citation

According to the traditional method of citation analysis, the statistics of each literature on the total citation frequency, found data maximum co-citation frequency was 9, so the author for each citation frequency (1-9) statistic of the hierarchical distribution of Co-citation respectively, see table 4.

Table 4. Frequency and Level of Co-citation of JOI Journals Published in 2007

<table>
<thead>
<tr>
<th>Frequency of co-citation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two times of sentence</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Sentence</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>33</td>
</tr>
<tr>
<td>Paragraph</td>
<td>163</td>
<td>235</td>
<td>52</td>
<td>14</td>
<td>20</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>515</td>
</tr>
<tr>
<td>Section</td>
<td>124</td>
<td>352</td>
<td>49</td>
<td>23</td>
<td>13</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>571</td>
</tr>
</tbody>
</table>
As can be seen from the table two, in the same literature more than 11 times a total of 2 references were cited in the sentence level, of which 8 were cited in accordance with the conventional method of co-citation analysis. In the same literature, more than two times of the total cited in the sentence level of the literature on, intuitive, is very similar, but in accordance with the traditional way of analysis of the results, the similarity is relatively weak. In addition, table 4 of the cited frequency was 1, 2 and 3 of the literature accounted for more than 90% of the total, while only about 1/3 of the total citation is in the layer, if the traditional method to calculate the total citation frequency measure co-citation strength, while greatly weakening the similarity between documents. In the co-citation analysis, the similarity of the literature is not only related to the co-citation frequency, but also with the co-citation level.

We use SPSS statistical analysis software to carry out contingency table analysis, and analyzes whether there is a certain correlation between the co-citation frequency and the co-citation level. Column variable contingency table for total citation frequency, variable as co-citation level, which is defined as "1 sentence 2 times", 2 "3" sentence "paragraph layer", 4 ", 5" chapters "in this layer, resulting in Table 5.

Table 5. Co-citation level and Co-citation Frequency Cross Contingency Table

<table>
<thead>
<tr>
<th>Co-citation Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Number</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Expect number</td>
<td>3.2</td>
<td>5.8</td>
<td>0.9</td>
<td>0.4</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>11.0</td>
</tr>
<tr>
<td>Co-citation level (%)</td>
<td>0.0%</td>
<td>72.7%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>18.2%</td>
<td>0.0%</td>
<td>9.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Co-citation frequency (%)</td>
<td>0.0%</td>
<td>0.9%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>4.9%</td>
<td>0.0%</td>
<td>7.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.6%</td>
</tr>
<tr>
<td>Total (%)</td>
<td>0.0%</td>
<td>0.5%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.6%</td>
</tr>
</tbody>
</table>

| 2 Number          | 6 | 6 | 3 | 4 | 2 | 4 | 3 | 3 | 2 | 33    |
| Expect number     | 9.7 | 17.5 | 2.7 | 1.1 | 0.8 | 0.3 | 0.3 | 0.3 | 0.3 | 33.0 |
| Co-citation level (%) | 18.2% | 18.2% | 9.1% | 12.1% | 6.1% | 12.1% | 9.1% | 9.1% | 6.1% | 100.0% |
| Co-citation frequency (%) | 1.2% | 0.7% | 2.2% | 6.8% | 4.9% | 28.6% | 21.4% | 17.6% | 13.3% | 1.9% |
| Total (%)         | 0.4% | 0.4% | 0.2% | 0.2% | 0.1% | 0.2% | 0.2% | 0.2% | 0.1% | 1.9% |

| 3 Number          | 163 | 235 | 52 | 14 | 20 | 8 | 8 | 8 | 7 | 515    |
| Expect number     | 151.6 | 272.8 | 42.1 | 17.9 | 12.4 | 4.2 | 4.2 | 5.2 | 4.5 | 515.0 |
| Co-citation level (%) | 31.7% | 45.6% | 10.1% | 2.7% | 3.9% | 1.6% | 1.6% | 1.6% | 1.4% | 100.0% |
| Co-citation frequency (%) | 32.6% | 26.1% | 37.4% | 23.7% | 48.8% | 57.1% | 57.1% | 47.1% | 46.7% | 30.3% |
| Total (%)         | 9.6% | 13.8% | 3.1% | 0.8% | 1.2% | 0.5% | 0.5% | 0.5% | 0.4% | 30.3% |

| 4 Number          | 124 | 352 | 49 | 23 | 13 | 2 | 0 | 4 | 4 | 571    |
| Expect number     | 168.0 | 302.5 | 46.7 | 19.8 | 13.8 | 4.7 | 4.7 | 5.7 | 5.0 | 571.0 |
| Co-citation level (%) | 21.7% | 61.6% | 8.6% | 4.0% | 2.3% | 0.4% | 0.0% | 0.7% | 0.7% | 100.0% |
| Co-citation frequency (%) | 24.8% | 39.1% | 35.3% | 39.0% | 31.7% | 14.3% | 0.0% | 23.5% | 26.7% | 33.6% |
| Total (%)         | 7.3% | 20.7% | 2.9% | 1.4% | 0.8% | 0.1% | 0.0% | 0.2% | 0.2% | 33.6% |

| 5 Number          | 207 | 299 | 35 | 18 | 4 | 0 | 2 | 2 | 2 | 569    |

772
After the table 5 is compiled, we use the method of chi square test to analyze whether there is a connection between the co-citation frequency and the co-citation level. In this paper, the original hypothesis of chi square test is that the co-citation frequency and the co-citation level are independent of each other. The chi square test results are shown in Table 6, and the cluster bar graph is shown in Figure 4.

### Table 6. Chi Square Test

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>234.539</td>
<td>32</td>
<td>0.000</td>
</tr>
<tr>
<td>Continuity Correction</td>
<td>163.304</td>
<td>32</td>
<td>0.000</td>
</tr>
<tr>
<td>Linear-by-Linear Association</td>
<td>69.011</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>1699</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Figure 4. Cluster Bar Chart](image)

Table 6 shows that the chi square statistic (Chi-Square Pearson) value is 234.539, the corresponding p value is 0. If the significance level alpha is set to 0.01, the probability of P chi square value is less than alpha, so we should reject the null hypothesis, that the co-citation and co-citation levels are related, and the linear by linear association (Linear-by-Linear Association) the probability of P is smaller than that of alpha, total cited frequency and co-citation level linear. From table 4, figure 4 shows that the co-citation level total citation frequency is mainly 2 and 1, with the increase of frequency, the level of co-citation presenting the possibility of a decreasing trend, the total citation frequency can be determined with the co-citation level co-citation is
negatively linear correlation.

4.3 Comparison of cluster analysis

According to the traditional co-citation analysis method and the position based co-citation analysis method, we establish the co-citation matrix of these 78 documents, and then introduce the matrix into SPSS software to compare the cluster analysis. Firstly, the clustering results of the two methods are obtained by using the system clustering method, as shown in Figure 5.

![Figure 5. Traditional (left) and Position-based (right) Co-citation Analysis](image)

According to figure two, the clustering results of the 5 methods are found to be more precise than the traditional methods based on the consistency of the clustering results, as shown in Figure 5. Secondly, the clustering effect index is introduced, and the clustering results of the two methods are compared and evaluated.

Results the best clustering should make the similarity class of internal documents as large as possible, the class similarity between documents is as small as possible, according to this standard,
and the author uses the clustering quality evaluation comparison of the two methods of clustering results (Han and Zhao, 2009). The indicators are defined as follows:

\[
S_{ter}(C) = \frac{\sum_{i=1}^{K} S_{tera}(C_i)}{K} \quad (1)
\]

\[
S_{tra}(C) = \frac{k}{1 + \sum_{j=1}^{k} \text{dist}(\text{Centroid}_j, \text{Centroid}_i^2)} \quad (2)
\]

\[
CF = \frac{2S_{tera}(C) / S_{tra}(C)}{S_{tera}(C) + 1 / S_{tra}(C)} \quad (3)
\]

Among them, clustering results within class similarity, poly class number, is a subset of the results in class; the similarity clustering results, is the center of the J class, is the center of all points. Therefore, the larger the value, the better the clustering effect.

In this study, a total of 78 articles are involved in clustering, at least can be clustered into 1 categories, up to 78 categories, the number of clusters is too small to lose the significance of clustering, according to the sample data to be classified. Figure 5 analysis can be found, re adjust the distance of 25 is a critical value, when the distance is slightly less than 25, two methods are clustered into 6 categories; when re adjust the distance of 15, based on the position of the traditional co-citation analysis and cluster number were 18, 16, two methods were 11 Article 8, article alone into classes, accounted for the total number of class 61% and 50%; when re adjust the distance of 10, based on the position of the traditional co-citation analysis and cluster number were 29, 25, the two methods were 18 articles, 16 articles and 62% separate categories, accounting for 64% the total number of classes; when re adjust the distance of 5, a single article into a single cluster ratio of more than 80%. In conclusion, the author thinks that 78 articles should not be more than 18 the number of clusters. Continue to use SPSS software, K-means cluster analysis, and calculate the number of fixed class in a number of two methods of clustering results of the quality of the results, in order to compare. Table 7 lists the 6 class to 18 class together, the corresponding position clustering results with the traditional co-citation analysis method based on quality.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>CF_{position}</th>
<th>CF_{tradition}</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.83</td>
<td>1.25</td>
</tr>
<tr>
<td>7</td>
<td>0.99</td>
<td>1.32</td>
</tr>
<tr>
<td>8</td>
<td>1.11</td>
<td>1.25</td>
</tr>
<tr>
<td>9</td>
<td>1.38</td>
<td>1.31</td>
</tr>
<tr>
<td>10</td>
<td>1.43</td>
<td>1.36</td>
</tr>
<tr>
<td>11</td>
<td>1.48</td>
<td>1.30</td>
</tr>
<tr>
<td>12</td>
<td>1.39</td>
<td>1.27</td>
</tr>
</tbody>
</table>

Table 7 analysis illustrate that when clustered into 11 classes, the position of the co-citation analysis method to achieve the best clustering result based on the position value is 1.48, the traditional value is only 1.30; when clustered into 10 classes, the traditional co-citation analysis method to achieve the best clustering effect, the traditional value is 1.36, and the position value 1.43; when the cluster number is less than 9, the traditional co-citation clustering analysis method
is better than the quality of the clustering results of co-citation method based on position of quality, but the two methods have not yet reached the best clustering results, clustering number is not desirable; when the cluster number is not less than 9, clustering the results of co-citation clustering results the position of quality is superior to the traditional co-citation analysis method based on quality. Therefore, with the traditional co-citation analysis method, the position of the co-citation analysis methods to enhance the co-citation clustering results based on the analysis of the relationship between the clustering (or more weak) more literature from the formation of small separation categories, the similarity of the internal documents (more), class the similarity between the literature (smaller). Based on the position of the co-citation clustering method for more accurate clustering of scientific literature, it is conducive to the in-depth study and evaluation.

5. Discussion and Conclusions

This paper constructs the position of co-citation analysis method based on the framework, the co-citation analysis is divided into four levels and the differences between the assignment respectively; secondly, in the empirical analysis through the analysis of the citation frequency of each chapter of the position information can be found according to the citation, Citation Distribution classification on the scientific literature; third, through the co-citation analysis hierarchical distribution, that accounts for about 66% of the total citation is at the sentence level, paragraphs and chapters layer, the traditional co-citation analysis results cover the non-article level effect; fourth, the relationship between the level and cited by analyzing references find the co-citation frequency and a negative linear correlation between the co-citation co-citation level; fifth, the use of SPSS to two kinds of co-citation matrix clustering comparison, found that the position is based on The cited analysis method improves the validity of co-citation clustering.

Although the research made some achievements, but there are two problems: first is that the data is imitated in sample size and coverage, the research conclusions in other disciplines are founded, has yet to be verified; second is lacking of co-citation distance assignment f, four layer co-citation represents co-citation strength gap is equal, still needs to confirm the specific method of calculation.

In general, the position of the co-citation analysis method compared with the traditional co-citation clustering analysis method has a better effect on, although not confirmed the rationality of valuation gap, but the new method can help to improve the co-citation analysis, based on the position of the visible co-citation analysis has a very important value and the significance the development of co-citation analysis.

Acknowledgments

This paper is supported by National Social Science Foundation in China (Grant No.16BTQ055)

References:


Liu Shengbo and Zhang Chunbo et al (2013). The Improvement of Co-Citation Analysis Based on the Citation Context and Citation Position. *Journal of The China Society for Scientific and Technical Information*, 32 (12): 1248-1256.


Zhao Rongying and Zeng Xianqin et al (2014). Citation in Full-text: The Development of Citation Analysis. *Library and Information Service*, 58 (9): 129-135.
Sleeping Beauty awakened by self-citation of a review: a case study of Judah Folkman hypothesis on angiogenesis.

Philippe Gorry¹ and Adil El Aichouchi²

¹ philippe.gorry@u-bordeaux.fr
GREThA UMR CNRS 5113, University of Bordeaux, (France).

² adil.elaiouchi@outlook.fr
SPH EA 4574, Bordeaux Montaigne University, (France).

Abstract
Judah Folkman is considered the father of angiogenesis research. However, his hypothesis on tumor angiogenesis met a lot of skepticism at the beginning. Scientific resistance has been described in sociology of science, and leads to delayed recognition of pioneer work. In bibliometrics, it is characterized by papers, called sleeping beauties, that do not achieve recognition in terms of citations until a few years after their original publication. The present paper explores to what extent the phenomenon of delayed recognition have affected Folkman’s scientific production and the citation life of his publications. Citation analysis show that Folkman’s landmark paper published in 1971 is a sleeping beauty. The scientometric analysis is associated with a qualitative analysis in order to shed a light on the reasons behind the delayed recognition, and the awakening of “Sleeping Beauty” by a “Prince”, attracting a lot of attention in terms of citations. Interestingly, the fact that Judah Folkman was one of the co-authors of the Prince paper challenges the common practice of excluding self-citations and review while conducting such bibliometric analyses.

Conference Topic
Studies on the level of individual scientists

Introduction
Judah Folkman’s scientific production
The field of angiogenesis research began in the late 1960s with an attempt to delineate the role of neovascularization in tumor growth (Folkman, 1963). Over the past 50 years, angiogenesis research has become a very active and broad field of biomedicine. This new field of biology was initiated and pioneered by the American medical scientist Judah Folkman (1933-2008), who is unanimously considered the father of angiogenesis (Ribatti, 2008). Folkman introduced the concept of “anti-angiogenesis” as a potential novel anticancer therapy and his discovery is considered as a paradigm shift in cancer therapy (Cao, 2008). As a result, more than 1.2 million patients worldwide are now receiving antiangiogenic therapy (Bischoff, 2008).
Folkman received more than 150 awards from 11 nations. He was also elected to the National Academy of Sciences, to the Institute of Medicine and appointed to the President’s Cancer Advisory Board (Klagsbrun, 2008). Yet, during the first 20 years of his career, his ideas were met with a lot of skepticism, criticism and rejection (Folkman, 2008a). Folkman liked to reflect on the resistance he faced from his pairs by saying: “If your idea succeeds, everybody says you’re persistent, if it doesn’t succeed you’re stubborn” (Oranski, 2008).
Delayed recognition in scientific literature

The early rejection of Folkman’s ideas by the scientific community leads to the hypothesis that his work suffered from “delayed recognition”. Delayed recognition is a phenomenon where papers do not achieve recognition in terms of citations until a few years after their publication (Garfield, 1989a, 1989b, 1990; Glänzel, 2003; Van Calster, 2012). It also refers to terms such as “resisted discoveries” (Barber, 1961), “premature discoveries” (Merton, 1988), “late-bloomers” (Stent, 1972), “Mendel syndrome” (Costas, 2011). In scientometrics, it is called “sleeping beauty” (SB) (Van Raan, 2004), a publication that goes unnoticed for a long time, and then, awakened suddenly by a ‘prince’ (PR), attracting a lot of attention in terms of citations.

Quantitative criteria for studying delayed recognition can be summarized as being of three kinds: average-based criteria, quartile-based criteria and parameter-free criteria. One example of average-based criteria is van Raan’s (2004). Van Raan defined three variables for SB (a) depth of sleep (cs), that is, average citations per year received in the sleeping period since publication, with cs< = 1 standing for deep sleep and 1<cs< =2 for less deep sleep; (b) length of sleep (ns), that is, duration of sleeping period, often lasting between 5 and 10 years; and (c) awake intensity (Cw), amounting for instance to 20 citations per year over a period of 4 years. The quartile-based criteria were presented by Costas and colleagues’ measure (2010). They are able to describe three categories of publications by identifying the “Year 50%” as the year in which an article received for the first time at least 50% of its citations. SB are papers that have not received 50% of their citations when 75% of other documents in their fields have already received 50% of their citations.

A parameter-free index was proposed by Ke et al. (2015): “The Beauty coefficient (B)”. It quantifies the extent to which a paper could be considered a SB by adding up differentials between the citation curve of the publication and a reference line calculated between the year of publication and the year of maximum citations. Applying their criteria to their database, they consider as “top sleeping beauties” the 1000 papers with the highest “Beauty coefficient”, which corresponds to papers with B>=317.93.

Like in the fairy tale, a “Sleeping Beauty” must be awoken by a Prince. The Prince is usually the author of the first citing paper, but Du and Wu propose that candidates for "prince" should fulfill additional criteria such as: (i) the PR should be published at the time when the SB began to be highly cited; (ii) the PR should be highly cited papers themselves; and (iii) PR should be co-cited with the SB (Du, 2016).

The quantitative criteria that scientometrics offer for studying delayed recognition can be very useful to understand the dynamics of scientific change. However, such analysis must be examined critically by qualitative approaches such as historical or sociological analyses.

Aim

Sleeping beauties are relatively rare (<0.1%) (Ke, 2015) but they have been identified in numerous medical or research fields such as chemistry (Van Raan, 2004), metallurgy (Van Raan, 2004), ophthalmology (Ohba, 2012), pediatrics (Zavrsnik, 2016), psychology (Ho, 2017), or radiology (Gorry, 2016), and popularized among the wider scientific community (Cressey, 2015). The reasons for the SB pattern of citations may be linked to a paradigm shift in the research field (Van Raan, 2004) or social recognition through a Nobel Prize, for example (Du, 2016). However, the explanations for sleeping beauties are under explored.

The study of delayed recognition is necessary to understand scientific knowledge better. The present work aims to study the extent to which the phenomenon of delayed recognition affected Folkman’s body of work. We analyzed Folkman’s scientific production and the citation life of his publications looking for sleeping beauties. Then, we attempted to identify the relevant
prince, and tried to understand the reason for delayed recognition and the awakening mechanisms.

**Methods**

A collection of 510 publications authored by Judah Folkman were extracted with metadata, as well as a corpus of 116,703 citations from Scopus® database through 31 December 2015. Descriptive statistics were conducted using the Scopus® built-in functions and exported in CSV format to Excel, which was employed for further calculation. Using Ke at al.’s criteria (2015), the “Beauty coefficient” (B) was calculated for all of Folkman’s papers in order to identify top sleeping beauties. B quantifies the extent to which a paper may be considered a SB. For a given paper, they defined $c_t$ as the number of citations received in the $t^{th}$ year after publication, and $t$ as the age of the paper. The paper receives its maximum number $c_{tm}$ of yearly citations at time $t_m$. Straight line, $L_t$ connects the points $(0, c_0)$ and $(t_m, c_{tm})$ in the time-citation plane. The beauty coefficient B is then calculated using the Ke et al.’s equation. Ke et al. offer also a simple method for identifying the awakening time $t_a$ which is defined as the time $t$ at which the distance $d_t$ between the point $(t, c_t)$ and the reference line $L_t$ reaches its maximum.

Following Costas et al. (2011), we also calculated the quartile-based criteria “Year 50%” for all articles of the same year of publication, organized them in ascending order in terms of their percentiles, and recorded those falling on percentiles 25 as “P25” and those falling on percentiles 75 as “P75”.

The SCImago® journal rank measure was directly retrieved from Scopus®. The numbers of co-citations between the SB paper and each candidate Prince paper were calculated using Gephi® an open-source network analysis software package (https://gephi.org/). Each citation is represented by a unique Scopus identifier (EID). The number of co-citations was equal to the number of citation duplicates detected by Gephi®.

Finally, all the bibliometrics analysis were complemented by a historical approach of the Folkman biography and history of angiogenesis using sources with criticism.

**Results**

*Judah Folkman’s scientific production*

Folkman published his first paper in 1953 while he was enrolled at Harvard Medical School in 1957. In 1967, he was appointed Surgeon-in-Chief at Children’s Hospital, and the year after Professor of Pediatric Surgery at Harvard Medical school, at the unprecedented age of 36. He stepped down from that position in 1981 in order to work exclusively on angiogenesis research. During the 41 years that he worked at the Boston Children Hospital, he was the author of 396 articles, 41 reviews, and some 76 monographs. He published on the average 9.658 papers/year with a maximum of 23 papers in 1998 (Fig.1; grey box; right y-axis: publications number).

All along his scientific career, he published in various disciplines (multidisciplinary, medicine, surgery, oncology, pathology, biochemistry) in more than 160 different journals. About one-quarter of his research publications was published in the best medical and biology journals (Table 1). He also collaborated with a large number of researchers (more than 150) during his prolific and long career. His main collaborators were his fellows or technician, Yueng Shing, Evelyn Flynn, Robert D’Amato, and Michael O’Reilly.

Folkman’s scientific work was rapidly recognized in the academic community (675 total citations in 1970 for 26 publications); until 2015, his work was cited more than 116,700 times (Fig. 1; dashed line: cumulative citations for all Folkman’s publications by year, left y-axis).

Regarding his main research interest, he published his first paper on tumor growth in 1963, and published in 1971 his first two paper untitled angiogenesis (Fig. 1: black box).
Table 1. Top journals where Folkman published the most.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Source Title</th>
<th>Publications number</th>
<th>SJR2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Proc. Natl. Acad. Sci. USA</td>
<td>26</td>
<td>6.863</td>
</tr>
<tr>
<td>2</td>
<td>Cancer Research</td>
<td>21</td>
<td>5.372</td>
</tr>
<tr>
<td>3</td>
<td>Journal of The National Cancer Institute</td>
<td>17</td>
<td>6.192</td>
</tr>
<tr>
<td>4</td>
<td>Science</td>
<td>16</td>
<td>13.217</td>
</tr>
<tr>
<td>5</td>
<td>Journal of Pediatric Surgery</td>
<td>15</td>
<td>0.802</td>
</tr>
<tr>
<td>6</td>
<td>Nature</td>
<td>14</td>
<td>21.936</td>
</tr>
<tr>
<td>7</td>
<td>American Journal of Pathology</td>
<td>13</td>
<td>2.653</td>
</tr>
<tr>
<td>8</td>
<td>Annals of Surgery</td>
<td>12</td>
<td>4.503</td>
</tr>
<tr>
<td>9</td>
<td>New England Journal of Medicine</td>
<td>12</td>
<td>14.619</td>
</tr>
</tbody>
</table>


delayed paper has a B coefficient well above the threshold (B>=317.93) used by Ke et al. (2015) unlike the following ones (Table 2). This paper entitled “Tumor Angiogenesis: Therapeutic implications” was a review which was published in 1971, in a top journal (99th percentile rank): the New England Journal of Medicine (Folkman, 1971). This landmark paper has averaged 139.53 citations per annum, amounting to 6,279 citations up to the 31st December 2015.

Figure 1. Folkman’s publications and citations.

A single sleeping beauty

The calculation of the “Beauty coefficient” using equation (1) for all of Folkman’s papers revealed the existence of only one top SB with a sleeping beauty coefficient B=1052.17. This delayed paper has a B coefficient well above the threshold (B>=317.93) used by Ke et al. (2015) unlike the following ones (Table 2). This paper entitled “Tumor Angiogenesis: Therapeutic implications” was a review which was published in 1971, in a top journal (99th percentile rank): the New England Journal of Medicine (Folkman, 1971). This landmark paper has averaged 139.53 citations per annum, amounting to 6,279 citations up to the 31st December 2015.
Table 2. Top 5 Folkman delayed papers ranked by decreasing B index.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Publication year</th>
<th>Journal</th>
<th>Citations number</th>
<th>B coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Folkman J.</td>
<td>1971</td>
<td>New England Journal of Medicine</td>
<td>6279</td>
<td>1052.17</td>
</tr>
<tr>
<td>Folkman J.</td>
<td>1972</td>
<td>Annals of Surgery</td>
<td>838</td>
<td>175.04</td>
</tr>
<tr>
<td>Folkman J. Hochberg M.</td>
<td>1973</td>
<td>Journal of Experimental Medicine</td>
<td>449</td>
<td>137.22</td>
</tr>
</tbody>
</table>

Figure 2. Citation history of Folkman’s sleeping beauty, candidate prince papers and angiogenesis publication trends.

During its first 23 years, however, the paper was cited only 113 times. This suddenly rose to 18 citations in 1995, 33 citations in 1996, and 65 citations in 1997 (Fig. 2; black line: number of citations per year for Sleeping Beauty, left y-axis). Since 1995, the paper has averaged 293.6 citations per annum. This clearly exemplifies a sleeping beauty, although it does not fit exactly into Van Raan’s definition. With its 4.7 citations per year during the sleeping period, the paper qualifies as a “light sleep” (Van Raan, 2004; 25); it slept for 23 years instead of five to ten. Also, the average number of citations over the four years following its awakening was 51 per
annum, which is a very high awake intensity by considering Van Raan’s definition (minimum of 20 citations per year for a 4-year period). In addition, using Costas et al.’s criteria (2011), Year50% = 39 (which corresponds to the year 2009), while P75 = 34.5. This means that Folkman’s paper is classified as a paper with delayed recognition.

**Identifying the Prince**

In order to determine the Prince paper, the year of the awakening was calculated using Ke et al.’s equation (2016) and was revealed to be 1994. Based on the awakening year, the investigation was narrowed down to the period of time between 1992 and 1996 in which the top publications citing Folkman’s main paper were identified. The Prince paper was likely to have been published in a top rank journal, to be among the first highly cited citing articles, and to share the largest number of co-citations with the SB paper during the awakening period 1993-97 (Table 3). The citation curve of Prince should match the citation awakening of the SB paper.

**Table 3. Top candidate Prince publications citing Folkman’s Sleeping Beauty around the awaking year (1993-1997).**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Authors</th>
<th>Source</th>
<th>Year</th>
<th>Citations number</th>
<th>Co-citations with SB</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Holmgren, L., O’Reilly, M.S., Folkman, J.</td>
<td>Nature Med.</td>
<td>1995</td>
<td>1389</td>
<td>244</td>
</tr>
</tbody>
</table>

The most cited article citing Folkman’s paper is a publication untitled “Angiostatin: A novel angiogenesis inhibitor that mediates the suppression of metastases by a Lewis lung carcinoma”, published in 1994 in Cell. The first author is Michael O’Reilly’s, a post-doctoral fellow at that time in Folkman’s laboratory, who was one of the co-authors (O’Reilly, 1994) (Fig. 2; black dotted line: candidate Prince paper #1, left y-axis). This article also shares the highest number of co-citations with the landmark paper (505 co-citations) (Table 3). Our Prince is therefore this paper with Michael O’Reilly. Beside the citation of Folkman’s main paper by the Prince, the awakening can be attributed by the discovery of his pioneering work by a whole research community. It is worthy of note that the awakening citations of Folkman’s paper matched with the annual rate of publications entitled or indexed (abstract, keywords) for the keyword angiogenesis (Fig. 2: black dashed line, right y-axis).

**Discussion**

*Delayed recognition of the angiogenesis-dependent tumor growth hypothesis*

Despite the fact that Judah Folkman is recognized today as the father of angiogenesis research, his landmark paper suffered from delayed recognition. Folkman’s SB paper first presented the hypothesis that “tumor growth is angiogenesis-dependent” (Folkman, 1971). In this rather
theoretical article, Folkman showed preliminary evidence that tumors could not enlarge beyond a microscopic size of 1-2 mm$^3$ without recruiting new capillary blood vessels. Folkman also introduced the term “angiogenesis” to mean the prevention of new vessel sprouts from being recruited by a tumor, reported the isolation of a tumor angiogenesis factor (TAF) and speculated that antiangiogenesis may provide a form of cancer therapy with an antibody against TAF. Many skeptics challenged Folkman’s hypothesis (Kerbel, 2000), and many scientists argued that the search for an angiogenesis inhibitor was a “fruitless exercise” (Folkman, 2008b). However, the 1980s witnessed the discovery of the first molecules that mediated angiogenesis such as acidic and basic fibroblast growth factor (Shing, 1984) and vascular endothelial growth factor (Ferrara, 1989). In addition, the first angiogenesis inhibitor, interferon α, was discovered by Brouty-Boyé & Zetter (1980). By 1988, daily low dose interferon α was successfully used to treat cancer. All those experimental and clinical advances throughout the 80s were not sufficient to trigger the awakening of Folkman’s paper. It was until the mid-90s that it received the attention and recognition it deserved.

The Prince’s kiss of life
Since the late 1970s, Folkman’s laboratory conducted a long-term effort to prove the existence of angiogenesis inhibitors. This effort was fruitful in that the laboratory reported eleven molecules between 1980 and 2005 (Folkman, 2008). In 1991, Michael O’Reilly came to Folkman’s laboratory as a postdoctoral fellow (Folkman, 1996). His mission was to attempt to uncover the mechanism of suppression of metastatic growth by a primary tumor. Folkman developed an alternative hypothesis after reading Noel Bouck’s report in 1989 that the emergence of tumor angiogenesis was the result of a shift in balance between positive and negative regulators of angiogenesis in a tumor (Rastinejad, 1989). All these findings suggested to Folkman that a primary tumor might suppress growth of its distant metastases by releasing an angiogenesis inhibitor into the circulation.

Indeed, the discovery of angio statin came as a result of an attempt to test this hypothesis (O’Reilly, 1994). The discovery and publication of angio statin in 1994 resulted in a fundamental evolution in the field of angiogenesis (Soff, 2000). The characteristics of Angio statin listed above made the scientific community finally appreciate the real importance and relevance of endogenous angiogenesis inhibitors (Folkman, 2004), and consequently recognize the major importance of Folkman’s 1971 founding paper. This late recognition was immediately translated by the subsequent awakening of Folkman’s sleeping beauty.

Conclusion
This paper has attempted to apply methodology of citation analysis to explore if Judah Folkman’s scientific work suffered from delayed recognition, and sustain the author feeling about scientific resistance to his tumor angiogenesis hypothesis (Bikfalvi, 2016). It has shown that Folkman’s landmark paper is indeed a sleeping beauty based on the calculation of the B coefficient and according to Costas’ criteria. If, it does not fulfil Van Raan’s sleeping beauty publication criteria, it is due to the arbitrary thresholds on the sleeping period and on awakening intensity (Gorry, 2016; Li, 2016). However, the idea that papers with delayed recognition show the highest impact in their fields is supported by our case study (Costas, 2010).

However, Folkman’s SB paper is categorized as a “review” paper, and, there is no consensus so far to the inclusion or exclusion of this type of document in citation analysis. Although the definition of a review might vary across fields, journals and time, it usually presents an overview of recent research advances and highlights results inconsistencies. Folkman’s review hypothesizing that “tumor growth was angiogenesis-dependent led to a profound paradigm shift in oncology” (Cao, 2008). Therefore, it is tempting to refer to Kuhn’s scientific revolution
(Kuhn, 1970), and consider that Folkman’s sleeping beauty is probably the paper that initiated the field of angiogenesis research.

Folkman’s SB paper slept for more than twenty years until the discovery of angiostatin in 1994 (Soff, 2000). The most obvious explanation is that the scientific community was skeptical, waiting for sound experimental evidence before changing in paradigm, according to Popper’s cumulative process (Popper, 1961). Alternatively, scientific controversy might explain delayed recognition of the SB (Gorry, 2016). Folkman’s landmark paper might have suffered from controversies surrounding the financing of his lab and its partnership with Monsanto during that sleeping period (Hess, 2006). However, the controversy born of the difficulties of replicating Folkman’s results by other laboratories did not affected the awakening of his landmark paper by the Prince paper. Angiogenesis research started to spread to many laboratories and became a burgeoning field with hundreds of papers per year.

Interestingly, the fact that Judah Folkman was one of the co-authors of the Prince paper challenges the common practice of excluding self-citations when conducting such bibliometric analyses. Self-citation is defined as a citation in which the citing and the cited paper share at least one author in common. If self-citation is believed to sustain self-promotion, it may be justified by the cumulative nature of science, when authors refer to previous hypotheses, methods or results (Popper, 1961). The fact that Folkman kept citing his own paper after all those years demonstrates his phenomenal persistence and belief in the importance of his theory. It should certainly not be seen as means to artificially inflating citation rates in order to strengthen his position in the community (Glänzel, 2008). During the sleeping period, the NIH turned down Folkman’s grant proposal, and he was able to continue his research program by securing founding with private companies (Hess, 2006). Constancy and continuity in a research field are important components that ensure the development of new research subject area (Price, 1976).

**Acknowledgment**

This work was supported by an internship to Adil El Aichouchi granted by the Bordeaux’ University consortium on vascular aging (ATT-VIVA). We would like to thank Pascal Duris for his support and Andreas Bikfalvi for his comments on the biology of angiogenesis.

**References**


Reference behavior in the full text of scientific articles: A large-scale analysis

Kevin W. Boyack¹, Nees Jan van Eck², Giovanni Colavizza³, and Ludo Waltman²

¹ kboyack@mapofscience.com
SciTech Strategies, Inc. (USA)

² [ecknjpvan, waltmanlr] @cwts.leidenuniv.nl
Centre for Science and Technology Studies, Leiden University (The Netherlands)

³ giovanni.colavizza@epfl.ch
Digital Humanities Laboratory, École Polytechnique Fédérale de Lausanne (Switzerland)

Abstract
In this paper, we report distributions of in-text citations in full text articles from two large databases – the PubMed Central Open Access Subset and Elsevier journals – as functions of time and textual progression. Using over five million articles from these two databases, we show that the numbers of in-text citations (i.e., mentions) per reference have remained nearly constant over time, and that the number of mentions per reference decreases with reference age. The data also show that mentions are concentrated at the beginning and near the end of articles, but that the distribution changes with the number of mentions per reference. References ages peak on average at around the 30% point in an article, and decrease thereafter. Citation counts for references also peak at around the 30% point, suggesting that methods papers are more highly cited than other types of papers. Citation counts are also a function of mentions – those references mentioned only once are much more highly cited than those mentioned many times, which correlates with the notion of perfunctory references being mentioned less often than those references that are more central to a citing article.

Conference Topic
Journals, databases and electronic publications
Analysis of full text

Introduction
The increasing availability of full text from scientific articles in electronic format is a development with the potential to greatly impact citation analytics and the ability to accurately model the structure of science. Full text contains information not only on the exact locations of in-text citations within articles, but also the context in which a citation to previous work is made. Specific problems that can be addressed using full text data include classification of in-text citations by type and function, and improving measures of impact by weighting of citations based on polarity, typology, function, citing location, and perhaps other features as well. Weighting of citations also has the potential to impact our knowledge of the structure of science, in that document clustering (and the resulting maps) could be based on a more accurate measure of the relatedness between documents. These applications, although beyond the scope of this paper, motivate the current work, which characterizes the distributions of in-text citations (and associated features) in two large full text databases. A solid understanding of these distributions is required before advanced applications of full text data, such as those mentioned, can be most fruitfully pursued.

Study of in-text citations and related text from scientific documents using full text sources has a long history. Although both syntactic (the location of references) and semantic (the meaning of references) studies have been pursued (Ding, Liu, Guo, & Cronin, 2013), here we focus primarily on the syntactic aspect. The terminology used in previous studies of in-text citations is not consistent. Thus, to avoid confusion, we define our terminology here. A reference is an
item in the bibliography or reference list of a document. An in-text citation is a mention of a reference within the full text of a document. A reference can be mentioned one or more times in a document. Each mention is an in-text citation. We use the terms in-text citation and mention interchangeably in this article.

Over the years, many studies of the location or position of in-text citations have sought to identify the relative value of citations as a function of the position or the number of mentions. Implicit among many of these studies is the assumption that references that are more related to the citing article are the more valuable or essential references for that article. In reviewing prior work we focus on those studies that explicitly include mention location in their analysis. Early studies were necessarily done by hand with small datasets. In one of the first studies, Voos & Dagaev (1976) examined citations to a set of four highly cited articles, two from biology, one from medicine and one from physics. Despite their very small sample, their findings suggested that a) most mentions come from introduction sections, b) the location and the number of mentions – which early studies often referred to as op. cit. – were both important in determining the value of a citation, c) time was important, and d) different disciplines had different citation patterns.

Cano (1989) sought to study citation function and utility while also examining position. Using 344 references that were coded by function and utility by their authors, they found that references that were mentioned in a perfunctory and negational way were most often peripheral or of low utility. They also found that references classified as organic, conceptual, operational or evolutionary were more typically essential or of higher utility, and that mentions were more concentrated in the first 15% of an article. Maričić et al. (1998) examined citation contexts as well as locations using 11% of the mentions to a set of 357 articles, and suggested that references should be valued differently based on the section of the citing article in which they appear. They found references with relatively low “meaning” (or value) to be mentioned predominantly in the introduction, while those mentioned in other sections had higher meaning.

Bornmann & Daniel (2008) examined a set of 350 in-text citations to a set of articles written by grant applicants. Using the IMRaD (introduction, methods, results and discussion) structure, they found that while more mentions appeared in the introduction and discussion sections of citing articles, methods and results sections were slightly enriched with mentions to articles with higher citation counts. In perhaps the most detailed comparison with ground truth data available, Tang & Safer (2008) surveyed authors of 49 articles in biology and 50 articles in psychology who assessed the mentions in their articles for importance, reason for citation, and relationship to the cited author. They found that reference importance increased with numbers of mentions and more detailed discussion. In addition, the authors considered references mentioned in the methods and results sections to be most important, while those mentioned in the introduction section only were less important than those in other sections.

Several recent works have also focused on the distribution of in-text citations within full text. Hu, Chen & Liu (2013) counted the number of mentions per section for 350 articles from Journal of Informetrics, finding that half of the mentions are in the first 30% of the text. They graphically represented distributions of mentions for articles with four, five and six sections. The six-section representation showed the greatest differentiation in the number of mentions per section, with counts decreasing (per kilo-words) through the first four sections, then increasing for the fifth section, and decreasing again for the last section. Ding et al. (2013) also counted mentions by section for 866 JASIST articles, finding that the literature review
section had the largest numbers, and that the most highly cited articles were referenced in the introduction and literature review sections.

The largest study of in-text citation distributions, and the most similar to our work, is the recent study by Bertin et al. (2016), who showed distributions of mentions and reference ages as a function of text progression from 45,000 scientific articles published in PLOS journals. Given that the PLOS documents have relatively consistent styles by journal and have very well tagged section headers, they were able to characterize these distributions in terms of the IMRaD document style. Specific documents that have sections not in IMRaD order were reordered to IMRaD order, leading to the observation that mentions are most highly concentrated in the introductions of articles, followed by the discussion section. They also showed that reference ages (i.e., time between the publication years of a citing publication and a cited publication) are highest at the beginning of the introduction and in the methods section, decreasing at the end of the introduction and in the results and discussion sections. The fact that these distributions are very similar across PLOS journals led them to characterize these as “invariant” distributions.

Finally, Zhao, Cappello & Johnston (2017) coded 1473 in-text citations from 14 JASIST articles by citation function, using three categories that were considered influential and two categories that were considered non-essential. To limit references to those that are influential for the purposes of citation analysis, they suggest that “removing all citation occurrences in the Background and Literature Review sections and uni-citations in the Introduction section appears to provide a good balance between filtration and error rates.”

In summary, regarding distributions of in-text citations, there is a consensus among previous studies that mentions tend to be more concentrated at the beginnings (e.g., introduction and related work) and endings (e.g., discussion and conclusion) of articles than in the middle sections. There is also a rough consensus that references that are mentioned outside the introductory sections are perhaps the most valuable.

Our work expands specifically on Bertin et al. (2016) by examining in-text citation distributions for two much larger full text datasets – the PubMed Central (PMC) Open Access subset and a large portion of the Elsevier full text corpus. Using these much larger and disciplinarily broader data, we will show that there are significant variations in the distributions that were not seen in the strongly biomedical PLOS corpus. We also report citation count distributions as a function of text progression. The paper proceeds as follows. We first review relevant literature and then describe our datasets and analysis methods. Results are then reported along with key observations. The paper concludes with a summary and suggestions for additional work.

Data and Methods
We have obtained access to two large sources of full text scientific articles that are available in machine readable form – the PubMed Central Open Access Subset (PMCOA) and full text from Elsevier (ELS) journals. Elsevier content is also available in their ScienceDirect product. PMCOA contains roughly 36% of the articles from PubMed Central, and is comprised of a) articles from open access journals and b) articles that are required to be made publicly available under the National Institutes of Health (NIH) public access policy and legislative mandates. Given publisher embargos, inclusion of the latter type of articles is typically delayed 12 months after publication. Nearly all PMCOA articles are also indexed in PubMed. ELS contains articles from nearly 3,000 Elsevier journals. Given that Elsevier is the largest
publisher of journals in the world, ELS is the single largest source of full text scientific articles currently available, and covers most scientific and technical fields.

Each full text source was obtained and processed independently, at different times, by different institutions, and using different code, as will be explained below. Despite these differences, the same basic steps were applied to each source: downloading, filtering, parsing, database creation, reference matching, and analysis.

**PubMed Central Open Access Subset**

The PubMed Central Open Access Subset was downloaded in XML format and processed by SciTech Strategies in October 2015. The subset included data through mid-2015 and contained 1,113,891 individual records, of which 945,279 had an associated PubMed ID (PMID) and at least one reference. Most of the records without PMID were conference abstracts that are not indexed in PubMed. We further limited the data to articles that were classified in PubMed as either a ‘journal article’ or a ‘review’, that were published in 1998 or later, and that had at least one reference with a reasonable reference year (defined as being between 1900 and 2015 and where the publication year was no earlier than one year prior to the reference year).

Each XML record was parsed such that individual sections, paragraphs and sentences were identified and numbered. While sections and paragraphs were delimited using XML tags, sentences were delimited and split using the NLTK (http://www.nltk.org/) pre-trained Punkt tokenizer for English. References, along with their bibliographic metadata (including PMID in many cases) were also extracted. In-text citations and their exact positions in the text (in terms of character offsets and text progression centiles within the article) were identified in the sentence level data using reference tags. Multiple in-text citations in the same bracket, whether using author/date or numbered formats, were given the same position in the text, regardless of which reference was listed first. Figure and table captions and footnotes were not considered. Parsed data were loaded into a MySQL database for further processing.

In addition to its centile position within the text, other data were added to each mention including the number of mentions (for the citing article and reference combination), reference age, the CWTS Leiden Ranking field to which the citing article belongs, and the Scopus citation counts to the reference as of the publication year of the citing article. Citation counts were obtained in a two-step process. First, Scopus article IDs were identified for each reference where possible by 1) using the listed PMID for the reference and looking up the corresponding Scopus ID from our matching table, or 2) matching the reference metadata to publication metadata from Scopus. Scopus IDs were identified for nearly 90% of the mentions. Citation counts to each reference were then from our Scopus data tables and added to the data. We note that citation counts are incomplete for references published prior to 1996 since Scopus records only contain references from 1996 onward. Leiden Ranking fields were added using a DOI lookup table provided by CWTS; only 593,224 articles could be matched to a field given the incompleteness of DOI data.

**Elsevier Full Text**

The Elsevier full text data were obtained in January 2017 and thus included a nearly complete 2016 publication year. Data were downloaded and filtered by CWTS using the following steps. First, the CrossRef REST API was used to identify all publications in Elsevier journals, numbering 8,437,487. Other types of publications, for instance publications in book series or conference proceedings, were not considered. Second, the Elsevier ScienceDirect API (Article
Retrieval API) was used to download these publications in XML format. Downloading was possible only for the 7,862,859 publications to which Leiden University has access via subscription. For some of these publications, the XML included only metadata rather than full text. Publications without full text were discarded, leaving 6,179,750 XML records. In particular, all publications that appeared before 1998 were discarded, because for almost all of these publications XML formatted full text was not available. Finally, the data were limited to those records that were English-language publications specifically labeled as full-length article, short communication, or review article, leaving a total of 4,821,774 full text records for analysis.

Each XML full text record was parsed to create sections, paragraphs, and sentences. This is important to properly locate each in-text citation. Sections and paragraphs were identified directly using XML tags. Only main sections were taken into account while subsections were ignored. Sentences could not be identified using XML tags. To identify sentences, we used a modified version of the sentence splitting algorithm provided in the BreakIterator class in the Java API. In-text citations and their exact positions in the text (in terms of character offsets and text progression centiles within the article) were identified in the sentence level data using the CWTS parsing algorithms. In addition, we also identified in-text citations occurring at various special locations in the full text of a publication, such as in footnotes and in the captions of tables and figures. Given that these mentions may perform a somewhat different function than mentions in the text body, we have not included them in the analysis that follows.

Publications in the ELS database were matched with publications in the Web of Science database. Publications were first matched based on DOI. If no DOI-based match was obtained, publications were matched based on the combination of the last name with first initial of the first author, publication year, volume number, and first page number, which together form a relatively unique key for each article. A match was required for all four fields. Of the publications retained for analysis, 4,503,790 could be matched with a publication in the Web of Science database. References in the full text publications were also linked to the five broad fields of science distinguished in the 2016 edition of the CWTS Leiden Ranking. Publications from 2016 were excluded in this step since the 2016 Leiden Ranking includes only publications through 2015. The number of publications that could be linked to one of the five fields is 3,892,083.

Descriptive Statistics
Numbers of articles, references and in-text citations for both datasets are given in Table 1, along with other characteristics of the data. Roughly 3% of the references were not included in the PMCOA analysis because they were missing a reference year or did not have a reasonable reference year. Although the average article in the ELS dataset is somewhat smaller than the average article in the PMCOA dataset (e.g., in numbers of paragraphs, sentences, characters, etc.), the average numbers of mentions per reference (1.573 and 1.572 for PMCOA and ELS, respectively) and percentage of sentences with mentions (19.91% and 19.77%) are very similar. The most substantial difference shown in Table 1 is in the numbers
of sections; this difference is explained by the fact that major sections and subsections were counted for PMCOA while only major sections were counted for ELS.

Table 1. Characteristics of the two full text databases.

<table>
<thead>
<tr>
<th>Table</th>
<th>PMCOA</th>
<th>ELS</th>
</tr>
</thead>
<tbody>
<tr>
<td># Articles</td>
<td>884,557</td>
<td>4,821,774</td>
</tr>
<tr>
<td># References</td>
<td>34,746,187</td>
<td>175,156,040</td>
</tr>
<tr>
<td># In-text citations (mentions)</td>
<td>54,649,985</td>
<td>275,337,977</td>
</tr>
<tr>
<td># Mentions w/citation counts</td>
<td>48,834,690</td>
<td>189,482,219</td>
</tr>
<tr>
<td>Avg # sections</td>
<td>12.27</td>
<td>5.70</td>
</tr>
<tr>
<td>Avg # paragraphs</td>
<td>39.01</td>
<td>32.59</td>
</tr>
<tr>
<td>Avg # sentences</td>
<td>179.19</td>
<td>152.45</td>
</tr>
<tr>
<td>Avg # sentences w/mentions</td>
<td>35.67</td>
<td>30.14</td>
</tr>
<tr>
<td>Avg # characters</td>
<td>27,582</td>
<td>24,588</td>
</tr>
<tr>
<td>Avg # references</td>
<td>39.29</td>
<td>36.33</td>
</tr>
<tr>
<td>Avg # locations w/mentions</td>
<td>45.04</td>
<td>36.72</td>
</tr>
<tr>
<td>Avg # in-text references</td>
<td>61.78</td>
<td>57.10</td>
</tr>
</tbody>
</table>

Results

In-text citations from full text articles can be characterized in a number of ways. For this paper, we choose to characterize distributions in terms of time and text progression.

Temporal Distributions

PMCOA and ELS have very different temporal distributions. PMCOA contains less than 1% of the articles indexed in PubMed prior to 2004, but has rapidly expanded since then to where it contains 19% of PubMed articles in 2014 (see Figure 1). In contrast, ELS contained over 100,000 articles in 1998, and while the number of full text articles has grown consistently, coverage (as compared with the total number of articles and reviews in Scopus) has stabilized at roughly 18% since 2004. Thus, PMCOA is highly skewed to recent content, while ELS contains a roughly consistent slice of the scientific literature over the past 12 years. Note that while we had fewer PMCOA articles in 2015 than in 2014 due to the timing of our data acquisition, full year values are shown for 2015 and 2016 (based on a recent query) as the dashed line in Figure 1.

Average numbers of references and mentions per article have increased in both databases over the years; however, the increase has been more pronounced for ELS than for PMCOA (see Figure 2). Despite the absolute differences, the average numbers of mentions per reference are very similar for the two databases and are nearly constant over time, decreasing only very slightly from 1.59 in 2001 to 1.57 in 2015. This constancy suggests that, despite changes in numbers of articles and references, the way in which references are mentioned has not changed recently.

A different type of temporal distribution can be examined by basing the data on citation interval (i.e., reference age). Figure 3 shows average numbers of references, mentions, and mentions per reference as a function of citation interval. Peak values occur for references and mentions for two-year old references; numbers then decrease steadily with increasing age. Most interestingly, the youngest references are mentioned more often than older references. While references aged two years or less are mentioned on average over 1.7 times, references older than 20 years are mentioned only 1.39 and 1.44 times each (PMCOA and ELS,
respectively). The differences between the tails in the PMCOA and ELS data likely stem from the fact that ELS covers all fields while PMCOA is focused on biomedical literature. Field-level differences will be investigated in future work.

Figure 1. Full text coverage by year.

![Graph showing full text coverage by year.]

Figure 2. Average numbers of references, mentions, and mentions per reference by year.

![Graph showing average numbers by year.]

Distributions by Text Progression

While several other studies have aimed to characterize distributions of in-text citations in terms of the IMRaD structure, our study simply characterizes these distributions as a function of text progression. We did not assign mentions to sections due to the lack of uniformity in section naming and ordering across journals. For example, while the ordering implied by IMRaD likely holds for a significant share of all journals, for other journals it is conventional for the methods section to be at the end of the article rather than after the introduction.
Figure 3. Average numbers of references, mentions, and mentions per reference by reference age. Bubble areas (right) denote relative fractions of mentions.

Figure 4 shows the distribution of mentions for three databases as a function of text progression using five-percentile bins. The reference distribution for the 66,227 articles in PLOS journals through 2012 that are available in PMCOA, approximating the results of Bertin et al. (2016) for the PLOS corpus, is included for comparison purposes. The three databases are similar in that the percentage of mentions for each is highest at the beginning of the paper (typically corresponding to the introduction), followed by a decay to a level that is relatively flat at around 3.5% of the references in each 5% bin. Each curve also then increases to a secondary peak nearer the end of the paper before tapering off to a very low level at the end.

In addition to these similarities, there are also subtle differences. PMCOA and PLOS have similar characteristics through the first 40% of articles, but the secondary peak occurs earlier (70% vs. 80%) and is a bit lower and broader for the PLOS curve than for the PMCOA curve.
despite the fact that both databases are primarily biomedical in nature. Bertin et al. (2016) showed that PLOS journals with the discussion section at the end have their secondary peak at the 90% point, while those with the methods section at the end have their secondary peak at the 60% point. PLOS ONE, with roughly half of each type, had a much shallower and broader peak at around the 70% point. The fact that the secondary peak in the PMCOA curve is shifted to the right of that in the PLOS curve suggests that PMCOA has proportionally more articles with their discussion section at the end. The ELS curve shows even more significant differences. The decay after the introduction is more gradual than that of the PMCOA curve and the secondary peak is much lower and occurs later, at the 85% point. These differences could easily be due to field-level differences. However, verification of this hypothesis will need to wait for future analysis. Regardless of the reasons for the differences, these results show that the distribution of in-text citations in scientific articles are perhaps not quite as invariant as previously reported.

Figure 5 shows distributions of mentions for the PMCOA and ELS databases as a function of number of mentions. The patterns are similar for both databases; the distributions of mentions flatten as the number of mentions increases. References mentioned only once are more prominent in the introduction sections of articles than are references mentioned many times.

We have also investigated citation interval as a function of text progression. Figure 6 shows results for PMCOA, ELS, and the PLOS subset. The three curves are similar in nature, with a slight decrease in reference age immediately after the first five-percentile bin, followed by an increase to a peak value and then a gentle decay to the end of the article. PMCOA reference ages are roughly one year older than those for PLOS articles. Given that both databases are largely biomedical, this suggests that the immediacy of PLOS articles is somewhat higher than those of biomedical articles in general. The ELS curve varies between one and two years higher than the PMCOA curve. Once again, we suspect that this difference is due to field-level effects.
Figure 6. Distribution of citation intervals (reference ages) associated with mentions for three databases.

Figure 7 shows distributions of citation intervals of mentions for the PMCOA and ELS databases as a function of number of mentions. The patterns are similar for both databases; the reference ages decrease as the number of mentions increases. References mentioned only once are substantially older than those mentioned many times, and are particularly so at the peak around the 25-35% point, likely corresponding in most cases to methods sections. This correlates well with the results of Figure 3, which showed that younger papers are mentioned more than older papers, but also shows that this holds true in all parts of an article.

Figure 7. Distribution of citation intervals associated with mentions as a function of number of mentions for PMCOA (left) and ELS (right).

Finally, we show the average citation counts for references mentioned as a function of text progression. Figure 8 shows that the PMCOA and ELS curves are similar in nature. References cited at the beginning of the article are more highly cited than those appearing later in the introduction. Citation counts then rise substantially to a peak at the 30% point, followed by a decrease, and then another increase toward the end of the paper. This final
increase is much more pronounced for PMCOA than for ELS, and may be specific to biomedicine.

![Figure 8. Distribution of citation counts associated with mentions for PMCOA and ELS (left), and average citation counts of references as a function of number of mentions (right). Bubble areas denote relative fractions of mentions (left) or references (right).](image)

Interestingly, the citation counts for references in PMCOA articles are much higher than those in ELS articles. We suspect that this difference is due to three reasons, but cannot do more than conjecture at this time. First, articles in biomedicine are often more highly cited than those in science as a whole due to citation cultures and database coverages. Second, PMCOA is mostly comprised of papers funded by NIH (due to the NIH OA mandate), and NIH funded papers have substantially higher counts than the typical biomedical paper (Boyack & Jordan, 2011). Even though we are measuring the citation counts of references rather than the citing papers, it is not unreasonable to suspect that articles that are highly cited also preferentially cite other highly cited articles. Third, Scopus citation counts are higher than citation counts from the Web of Science (WoS) for the same paper since Scopus coverage is broader than WoS coverage.

Figure 8 (right) also shows that the most highly cited references are typically mentioned less often than references that are less highly cited. Average citation counts for references mentioned only once are three times as high as for references mentioned eight or more times. References mentioned only once are also likely to be accompanied by less explanation than those mentioned multiple times (Zhao et al., 2017). This suggests that references mentioned only once are more likely to be perfunctory than those mentioned multiple times.

**Conclusions**

In this paper, we have analysed the in-text citations from over five million articles from two large databases – the PubMed Central Open Access Subset and Elsevier journals. Using these two databases, we have shown that the numbers of mentions per reference have remained nearly constant over time, and that the number of mentions per reference decreases with reference age. The data also show that mentions are concentrated at the beginning and near the end of articles, but that the distribution changes with the number of mentions per reference. References ages peak on average at around the 30% point in an article, and decrease thereafter. Citation counts for references also peak at around the 30% point, suggesting that papers cited in methods sections (which are typically other methods papers) are more highly cited than other papers. This contradicts the finding of Ding et al. (2013) that the most highly cited papers are mentioned in the introduction and literature sections. Citation counts are also a function of mentions – references mentioned only once are much more
highly cited than references mentioned many times, which correlates with the notion of perfunctory references being mentioned less often than those references that are more central to a citing article.

Differences exist between distributions from the PMCOA and ELS databases that cannot yet be explained. We suspect that these differences are primarily due to field-level differences. Immediate future work will thus focus on analysing these data using a field-level classification to identify the nature and magnitude of these differences, if any exist.

Acknowledgments
We thank Mike Patek for extraction and fielding of the full text from PubMed Central. Colavizza is funded by Swiss National Fund grant number P1ELP2_168489.

References

Chen Yue¹  Wang Zhiqi²  Tan Jianguo³  Liu Zeyuan⁴

¹chenyuedlut@163.com; ²zhiqi_wang90@126.com; ³tanjianguo0312@163.com; ⁴liuzy@dlut.edu.cn
WISE Lab, School of Public Administration and Law, Dalian University of Technology, Dalian, 116085 (China)

Abstract
This paper explores the position of Preprint in scholarly communication through bibliometric and empirical analysis of the arXiv-deposited papers in Information science & library science field and Robotics field, and focuses on four aspects, including “increasing trend”, “time gap”, “citation advantage” and “citation trend” in these two fields to demonstrate the effect of Preprint. The results show that the amount of preprints is increasing, most authors prefer to publish a pre-print in a journal or submit a post-preprint to arXiv within two years, the preprints have an obvious citation advantage compared with the papers only published in journals, and the citation increases rapidly after submitting the post-prints to arXiv.

Conference Topic
Journals, databases and electronic publications; Science communication

Introduction
Preprint, or Open Access (OA) shows an increasingly prosperous trend, and is widely supported by the government, research institutions and researchers all over the world. However, the progress of this emerging scholarly communication channel has been far slower in China compared with the world because of the academic evaluation and the academic journal issue system in China. Any promising thing will eventually break through the obstacles and move forward, the first Chinese preprint system, ChinaXiv was established on June in 2016 eventually. There emerged many preprint systems in 2016, such as engrXiv, SocArXiv, ChemRxiv and PsyArXiv. Especially, in the same year, the first preprint platform arXiv announced that 250-300 million dollars would be invested to promote its modernization and CrossRef announced to include the Preprint.

The emergence of a number of preprint platforms is challenging to the scientific journals. Scientific journals have been long dominant in the scholarly communication system through the printed papers by peer-review. The rejection and the delay due to lengthy journal review processes and manuscript backlogs, or no publish because of the payment, the block of access papers because of the economic burden as pay for the publisher, all of these problems in the scholarly communication systems would be partly solved by Preprint, due to the three main features of Preprint: “free academic publish”, “publish before print”, and “even publish no print”.

Bernal JD, the father of Science of Science, had promoted some insightful thought on scholarly communication as early as 1930s. He predicted that the rapid growth of science would lead to science journals actually abolished finally in his foundation work The Social Function of Science (1939), and expressed this opinion in The Transmission of Scientific Information: A User's Analysis (1959). He proposed to take each individual paper as the unit
of scholarly communication instead of journals, to promote scholarly communication directly using advanced electronic computers for scientific information collection, filtering and classification, to establish a central scientific literature system for scientists to transmit their research, as well as to retrieve and have access to all scientific literature. Scholarly communication is changing and will continue to change in future although the Bernal’s prediction is considered to be bold and exaggerated.

Preprints, as the name implies, initially referred to authors’ manuscripts submitted to an internet platform before published in journals. With the development of Preprint, such as arXiv, Coprints, RePEc, etc., the term “preprints” refers to all the papers deposited in preprint platforms at present, including both pre-prints (as submitted to a journal for peer review), working papers, post-prints (as a final copy of the peer-reviewed edited full text) or accepted manuscripts articles. Preprints—“temporary documents whose function is to bridge the time-gap created by publication delays” (Goldschmidt, 2002)—present a well-established mechanism for the exchange of scientific information.

Lawrence (2001) studied the correlation between articles freely available online (also called OA articles in some papers) firstly. He analysed 119,924 conference articles in computer science and related disciplines from DBLP (dblp.uni-trier.de), the results showed that freely available online articles were cited more often than ‘offline’ articles published in the same venue. Subsequently, many researches on the citation advantage of OA articles have emerged (Anderson, 2001; Craig, Plume & Mcveigh, 2007; Swan, 2010; Wagner, 2010; Ingwersen & Elleby, 2011; Gargouri & Lariviè re, 2012). The citation impact of Preprint varies by field. Brown (2003) explored the role of the Chemistry Preprint Server (CPS) in Chemical communication, and suggested that though it appeared unlikely that the CPS would replace the traditional chemistry journal in the near future, the CPS was utilized, appreciated, and valued by a wide variety of chemists to enrich their scholarly discourse and these extraordinary systems were likely to ultimately become the conventional scientific mode of information dissemination of the future. Brody & Harnad (2004) focused on the citation advantage analysis on pre-print manuscripts in physics and mathematics, and found that articles with a corresponding pre-print version deposited in arXiv would be more cited than those do not. Antelman (2004) studied the preprint in four subjects, Philosophy, Political Science, Electrical and Electronic Engineering and Mathematics, and found the proportion of preprints varied by disciplines from 17% to 69%, but across all these four disciplines, the preprints did have greater impact. Schwarz (2004), Metcalf (2005) and Henneken & Kurtz (2006) all found that papers with a matching pre-print version deposited in arXiv are both cited more than non-posted papers by examining the average citation rate of preprints in the Astrophysical Journal. In Solar Physics, there was also a significant citation advantage of preprints posted to arXiv or Montana State University (MSU) archive, which had citation rates 2.6 times and 1.7 times higher than the average of similar papers that are not posted as preprints respectively (Metcalf, 2006).

Some researchers explored the possible causes of citation impact of preprints. Kurtz (2005) proposed three possible explanations: “Early Access (EA)”, “Self-selection Preference (SP)” and “Open Access (OA)”, and found there is no OA effect in astrophysics. Hajjem (2005) suggested that the SP was unlikely the merely or mostly factor for citation advantage of preprints, because that though the proportion of preprints in physics was much higher (and
even approaches 100% in some subfields), the citation advantage of them was still obvious. Moed (2007) found that in the field of condensed matter physics, the EA and SP factors were the main causes for the citation advantages of arXiv-deposited articles, and which was also supported by Kurtz (2007). All the studies discussed above are only directly applicable to a certain discipline, but less universal across all research fields, as citation behaviour and author attitudes to Preprint are cultural factors that differ across different fields (Brown, 2003; Zitt, 2005; Alma & Brown, 2005). Many researchers have repeatedly confirmed the citation advantage of arXiv-deposited papers in many disciplines (Davis, 2007; Moed, 2007; Mine, 2009; Xin Shuai, 2012; Larivière et al., 2014), but few researchers do it in the field of LIS or Robotics, and analyse the “drive” factor for authors submitting a post-preprint to arXiv by comparing the citation trend before and after being submitted to arXiv.

In order to explore the position of Preprint in scholarly communication, this paper makes a bibliometric and empirical analysis of the arXiv-deposited papers in LIS field and Robotics field, and focuses on the four aspects: the usage trend of arXiv; the time gap between submission time and publication time; mean citation frequency of arXiv-deposited papers and WoS papers and their citation trend by year in these two fields.

**Data source**

Information science & library science (LIS) and Robotics are selected for this exploring research. In the LIS field, there are 43 journals listed in the Q1 and Q2 JIF quartiles in the LIS category in JCR (2015), and 20 of them are deposited in arXiv. According to the amount of arXiv e-prints (Tab.1), the top 6 journals (the total amount account for over 92%) are selected for our research.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Scientometrics</td>
<td>73</td>
<td>37.82%</td>
</tr>
<tr>
<td>2</td>
<td>Journal of Informetrics</td>
<td>52</td>
<td>64.77%</td>
</tr>
<tr>
<td>3</td>
<td>Journal of the American Society for Information Science and Technology</td>
<td>31</td>
<td>80.83%</td>
</tr>
<tr>
<td>4</td>
<td>International Journal of Geographical Information Science</td>
<td>12</td>
<td>87.05%</td>
</tr>
<tr>
<td>5</td>
<td>Research Evaluation</td>
<td>5</td>
<td>89.64%</td>
</tr>
<tr>
<td>6</td>
<td>Information Processing &amp; Management</td>
<td>5</td>
<td>92.23%</td>
</tr>
<tr>
<td>7</td>
<td>Learned Publishing</td>
<td>2</td>
<td>93.26%</td>
</tr>
<tr>
<td>8</td>
<td>Journal of Information Technology</td>
<td>1</td>
<td>93.78%</td>
</tr>
<tr>
<td>9</td>
<td>Journal of Computer-Mediated Communication</td>
<td>1</td>
<td>94.30%</td>
</tr>
<tr>
<td>10</td>
<td>Journal of The American Medical Informatics Association</td>
<td>1</td>
<td>94.82%</td>
</tr>
<tr>
<td>11</td>
<td>Information Systems Research</td>
<td>1</td>
<td>95.34%</td>
</tr>
<tr>
<td>12</td>
<td>Government Information Quarterly</td>
<td>1</td>
<td>95.85%</td>
</tr>
<tr>
<td>13</td>
<td>Journal of Knowledge Management</td>
<td>1</td>
<td>96.37%</td>
</tr>
<tr>
<td>14</td>
<td>Social Science Computer Review</td>
<td>1</td>
<td>96.89%</td>
</tr>
<tr>
<td>15</td>
<td>Information Society</td>
<td>1</td>
<td>97.41%</td>
</tr>
<tr>
<td>16</td>
<td>Online Information Review</td>
<td>1</td>
<td>97.93%</td>
</tr>
</tbody>
</table>
In the Robotics field, the journals (proceedings) are selected to research is on the base of robotics articles because there is no such an exact category in JCR. Top 10 journals are listed according to the amount of articles based on 58,799 robotics articles, which are retrieved by the query “WC=ROBOTICS” in WoS (2010-2017) (Tab.2), and the top 5 journals (proceedings) with large amount of preprints in arXiv (the total amount account for over 95% of all) are finally selected.

Table 2. Top 10 journals in robotics field deposited in arXiv

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IEEE International Conference on Robotics and Automation ICRA</td>
<td>6176</td>
<td>96</td>
<td>10.62%</td>
</tr>
<tr>
<td>2</td>
<td>IEEE International Conference on Intelligent Robots and Systems</td>
<td>5246</td>
<td>42</td>
<td>15.27%</td>
</tr>
<tr>
<td>3</td>
<td>Lecture Notes in Computer Science</td>
<td>3759</td>
<td>641</td>
<td>86.17%</td>
</tr>
<tr>
<td>4</td>
<td>Lecture Notes in Artificial Intelligence</td>
<td>3238</td>
<td>54</td>
<td>92.15%</td>
</tr>
<tr>
<td>5</td>
<td>Advances In Intelligent Systems And Computing</td>
<td>1919</td>
<td>8</td>
<td>93.03%</td>
</tr>
<tr>
<td>6</td>
<td>IEEE Asme International Conference on Advanced Intelligent Mechatronics</td>
<td>1603</td>
<td>3</td>
<td>93.36%</td>
</tr>
<tr>
<td>7</td>
<td>International Journal Of Advanced Robotic Systems</td>
<td>1304</td>
<td>31</td>
<td>96.79%</td>
</tr>
<tr>
<td>8</td>
<td>Communications in Computer and Information Science</td>
<td>1299</td>
<td>15</td>
<td>98.45%</td>
</tr>
<tr>
<td>9</td>
<td>Robotics and Autonomous Systems</td>
<td>1085</td>
<td>12</td>
<td>99.78%</td>
</tr>
<tr>
<td>10</td>
<td>Journal of Intelligent Robotic Systems</td>
<td>1005</td>
<td>2</td>
<td>100.00%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>26634</td>
<td>904</td>
<td></td>
</tr>
</tbody>
</table>

Results and Analysis

The increasing trend of arXiv-deposited papers

The proportion of arXiv-deposited papers in WoS is 2.17% and 0.27% in LIS and Robotics field respectively, which is much lower than the proportion of 18.7% in the physics field (Gentil-Beccot, 2010) and almost 100% in high-energy physics (Brody, 2006). This difference is understandable, because arXiv initiated in physics field (Brody, 2006). The amount of the arXiv-deposited papers is increasing generally in LIS and Robotics fields (Fig.1). The figure shows that the history of arXiv-deposited papers in Robotics is longer than in LIS, and the increasing trend of arXix-deposited papers in Robotics is more obvious than in LIS.
**Figure 1. The trend of submission time of arXiv-deposited papers**

*Time gap between arXiv submission year and journal publication year*

Most papers are submitted to arXiv and printed in journal in the same year in the two fields (Fig. 2). The preprints are divided into pre-prints and post-prints according to the order of submitted to arXiv and published in journal. In Fig. 2, the X-axis means the time gap between arXiv submission year and journal publication year and the Y-axis means the amount of Preprint papers. The result shows that the proportion of pre-prints and post-prints is 57.3% and 36.5% respectively in the LIS field and 47.4% and 41.4% respectively in the Robotics field. Most authors prefer to publish a pre-print in a journal or submit a post-preprint to arXiv within two years.

**Figure 2. Distribution of the time gap between arXiv submission and publication year**

a. the LIS field

b. the Robotics field
It should be noted that the amount of the preprints maybe underestimated because some article titles changed significantly enough to not be discoverable between their preprints and journal papers. In our research, the same articles are identified by matching on title, key words and abstract technically.

**Citation advantage of arXiv-deposited papers**

The potential impact advantage of the preprints has been a part of the wider Open Access debate (Harnad, 2004; Antelman, 2004; Metcalfe, 2004). The citation frequency per paper for arXiv-deposited papers and non-arXiv-deposited papers in WoS is counted respectively in LIS and Robotics field (Fig.3), which shows that the arXiv-deposited papers have the absolute citation advantage comparative to the non-arXiv-deposited papers. The citation advantage of arXiv–deposited papers in each journal is showed in Tab.3.

![Figure 3. Citation frequency per paper for arXiv-deposited and non-arXiv-deposited papers in WoS](image)

**Table 3. The citation advantage of arXiv–deposited papers in each journal**

<table>
<thead>
<tr>
<th>No.</th>
<th>Journal</th>
<th>Mean citation</th>
<th>Citation advantage*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>arXiv Scientometrics</td>
<td>22.208</td>
<td>112.81%</td>
</tr>
<tr>
<td></td>
<td>non-arXiv Scientific Information Systems and Technology</td>
<td>10.436</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>arXiv Journal of Information Science and Technology</td>
<td>12.255</td>
<td>1.50%</td>
</tr>
<tr>
<td></td>
<td>non-arXiv Journal of Information Science and Technology</td>
<td>12.074</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>arXiv Journal of The American Society for Information Science and Technology</td>
<td>46.226</td>
<td>249.24%</td>
</tr>
<tr>
<td></td>
<td>non-arXiv</td>
<td>13.236</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>arXiv International Journal of Geographical Information Science</td>
<td>16.363</td>
<td>66.64%</td>
</tr>
<tr>
<td></td>
<td>non-arXiv</td>
<td>9.820</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>arXiv Research Evaluation</td>
<td>14.200</td>
<td>95.90%</td>
</tr>
<tr>
<td></td>
<td>non-arXiv</td>
<td>7.249</td>
<td></td>
</tr>
</tbody>
</table>

In LIS field:
- arXiv-deposited: 22.208 (112.81%)
- non-arXiv: 10.436

In Robotics field:
- arXiv-deposited: 5.364 (18.45%)
- non-arXiv: 4.528

<table>
<thead>
<tr>
<th>No.</th>
<th>Journal</th>
<th>Mean citation</th>
<th>Citation advantage*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>arXiv IEEE International Conference on Robotics and Automation ICRA</td>
<td>5.364</td>
<td>18.45%</td>
</tr>
<tr>
<td></td>
<td>non-arXiv Scientific Information Systems and Technology</td>
<td>4.528</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>arXiv IEEE International Conference on Intelligent Robots and Systems</td>
<td>5.889</td>
<td>82.91%</td>
</tr>
<tr>
<td></td>
<td>non-arXiv</td>
<td>3.220</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>arXiv Lecture Notes in Computer Science</td>
<td>5.357</td>
<td>39.98%</td>
</tr>
<tr>
<td></td>
<td>non-arXiv</td>
<td>3.827</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>arXiv Lecture Notes in Artificial Intelligence</td>
<td>1.813</td>
<td>-28.69%</td>
</tr>
<tr>
<td></td>
<td>non-arXiv</td>
<td>2.542</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>arXiv International Journal of Advanced Robotic Systems</td>
<td>10.839</td>
<td>494.04%</td>
</tr>
<tr>
<td></td>
<td>non-arXiv</td>
<td>1.825</td>
<td></td>
</tr>
</tbody>
</table>
*Citation advantage= \((\text{citation frequency per arXiv-deposited} – \text{citation frequency per non-arXiv-deposited}) / \text{citation frequency per non-arXiv-deposited}\)"

**Citation trend of arXiv-deposited papers and WoS papers**

The citation trend of pre-prints, post-prints, highly-cited papers and WoS papers in the selected journals is showed respectively for LIS and Robotics field in Fig.4. Top 2% highly-cited LIS papers and top 1% highly-cited Robotics papers in each year are selected according to the proportion of preprints in these two fields. The X-axis represents the years after journal publication, and the Y-axis represents the citation frequency per paper in the corresponding year (the right Y-axis represents the mean citation frequency for the highly-cited papers).

![Citation trend in LIS and in Robotics](image)

**Figure 4.** Citation trend in LIS and in Robotics (wos-: all WoS papers in the selected journals; pre-: all pre-prints; post-: all post-prints; high-: highly-cited papers in the selected journals)
The citation trend of WoS papers shows upward in the first 3-4 years after journal publication, and then declines slowly in both the two fields. Comparatively, the citation trend of highly-cited papers shows a longer upward period and a quicker increase speed. The trends of pre-prints and post-prints are between the WoS papers and highly-cited papers, and fluctuating range of their citation curves is greater, which shows that preprints are more possible to be hot topics, which could attract more attention in a short time period. Comparing the citation trend of pre-prints with post-prints, in LIS, the pre-prints tend to be more cited than post-prints in general, while in Robotics, the situation is opposite in the first 3-5 years after journal publication, which could explain that why the proportion of pre-prints is larger than the post-prints in LIS, but opposite in Robotics to some extent.

It can be explained easier that authors submit a pre-print to arXiv before its journal publication, because pre-prints can give a longer period to be read and cited (Kurtz, 2005), and form citation advantage further. While, why the authors submit a post-print to arXiv? The relationship between the time period and the citation frequency of post-prints is benefit to this question. We set the arXiv submit time as the time division point, that is the ‘0’ on the X-axis, and investigate the citation trend of post-prints before and after they are submitted to arXiv respectively (Fig.5).

![Figure 5. The relationship between the citation and the interval time of post-prints](image)

In Figure 5, the X-axis represents the interval time between publishing a post-print in journal and submitting it to arXiv. For example, ‘-5’ means that the 5th years before submitting it to arXiv, and ‘1’ means the 1st year after submitting it to arXiv. The Y-axis represents the mean citation frequency per paper.

During the period that the authors haven’t submitted their journal papers to arXiv, the citation frequency of the post-prints increases at the rate of 0.186 and 0.063 in the field of LIS and Robotics respectively. However, it is clear that the citation frequency per paper increases faster at the rate of 0.253 and 0.457 after being submitted to arXiv respectively in the two fields. The results support that the arXiv platform could improve the citation frequency of the papers deposited in it, so some authors would still submit their journal papers to arXiv.
Conclusion and Discussion

There are different using patterns on the arXiv platform in the LIS field and Robotics field. The history of arXiv-deposited paper in Robotics is longer than in LIS, and the increasing trend of preprints in Robotics is more obvious than in LIS. The time gap distribution curves show that most authors prefer to publish a pre-print in a journal or submit a post-preprint to arXiv within two years.

In both fields, the citation advantage of arXiv-deposited papers is obvious. The citation trends of four kinds of papers—WoS papers, highly-cited papers, pre-prints and post-prints—show that the trends of pre-prints and post-prints are between the WoS papers and highly-cited papers, and fluctuating range of their citation curves is greater, indicating that preprints are more possible to be hot topics, which could attract more attention in a short time period.

The pre-prints tend to be more cited than post-prints in general in the LIS field, while the situation is opposite in the first 3-5 years after journal publication in the Robotics field, which could explain the proportion of pre-prints (57.3%) larger than the post-prints(36.5%) in LIS, but opposite in Robotics. The relationship between the time period and the citation frequency of post-prints represents that the mean citation frequency increases faster after being submitted to arXiv, so some authors would still submit their journal papers to arXiv.

All of the findings of this research indicate that the arXiv preprint archive has changed the scholarly communication patterns of the LIS and Robotics fields in a positive way. Though in these disciplines there is still a significant number of papers that are not on arXiv and the picture of ongoing changes in scholarly communication in a networked information environment is complex, the convenient, fast and high efficiency in scholarly communication through Preprint and citation advantage of preprints would prompt more authors to post preprints on the internet and these freely available papers would profit more researchers to have access to the newest research trend, which undoubtedly accelerates the process of scholarly communication and the knowledge innovation.

Our further study will extend to a wider perspective to explore the position of Preprint in scholarly communication in more disciplines, and other preprint archives such as E-LIS (the largest international open repository in the field of Library and Information Science) and ChinaXiv will also be the research objects.

Acknowledgements

The authors wish to thank Rousseau Ronald at IBW, University of Antwerp, and Department of Mathematics, KU Leuven, and Wolfgang Glänzel at the Department of MSI, Centre for R&D Monitoring (ECOOM), KU Leuven, for helpful comments on the manuscript.

This research was funded by the Project of Consultant and Research of Chinese Academy of Engineering under Gant ZX20160079 and the S&T Innovation Project of Liaoning Association for Science and Technology under Grant LNKX2016HBC01

References


Evaluation of Scholarly Impact from an Integrated Perspective of Three-Dimensional Citations: A Case Study of Gene Editing

Wang Feifei¹, Wang Xiaohan², Zheng Ran³, Liu Yang⁴

¹feifeiwang@bjut.edu.cn
Beijing University of Technology, Beijing (China)

²wangxiaohan0527@qq.com
Beijing University of Technology, Beijing (China)

³zhengranran112@hbpa.edu.cn
Hubei University of Police, Wuhan (China)

⁴170761977@qq.com
Chinese Academy of Agricultural Sciences, Beijing (China)

Abstract
On the basis of the integration of citation indexes, the traditional h-index, and the quantitative index of times cited, this paper proposes a comprehensive and comparative approach to analyze the influence of academic exchange, the diffusion of knowledge, and the contribution influence among authors. It integrates citation analysis, social network analysis, principal component analysis, entropy weight method, and Skyline operator from an integrated and comparative perspective of three-dimensional citation (author co-citation, author bibliographic coupling, and author cross-citation). Further, using the above idea, we conduct an exploratory research to find the scholars with comprehensive influence in the field of gene editing. The results were found to correspond well with the actual situation. Under the introduction of a weighted direct citation index, more important contributors to knowledge exchange and communication in this field were identified.

Conference Topic
Citation and co-citation analysis

Introduction
Research on scholars’ academic influence evaluation has been conducted for a long time. Scientific and effective evaluation of academic influence is essential both in terms of its basic function (e.g., motivating, guiding, and regulating) and specific application (e.g., performance appraisal, resource allocation, and talent introduction). The application of the information metrology theory and method in the academic influence evaluation, especially the use of citation analysis in traditional bibliometrics, has been a popular topic in the academic field. Especially since E. Garfield created the citation index and proposed the journal impact factor index, research workers and scientific research management personnel from various disciplines have begun directly or indirectly evaluating scholars’ academic influence by examining factors such as the number of papers published, cited frequency, h-index or its derivative index (Alonso, 2010), and journal impact factor. However, after the “San Francisco
Declaration on Research Assessment” and “Leiden manifesto” proposals, “make scientific research evaluation more scientific” has become a widely-discussed topic in recent years. The following are two core points:

1. Since impact factor is unsuitable for the evaluation of individual contributions, which or which group of quantitative indicators can replace it?

2. Since single indicator evaluation of academic influence is biased, how can various indicators be compared and integrated to make them more scientific and reasonable?

This paper aims to explore the optimization and comprehensive use of core quantitative indicators other than the impact factor and also focuses on the comparison and fusion of citation network indexes.

There are three main types of author associated relations based on the citation level: author co-citation, author bibliographic coupling, and author direct/cross citation. Each of the three kinds of citation analyses has unique applicability and emphasizes on revealing scholars’ academic association and reflecting their academic influence. The author co-citation network is mainly used to reveal core authors’ academic contribution; the author bibliographic coupling network mainly reveals the influence of the current research activity; and the author direct/cross-citation analysis indicates the inner link between authors while dynamically revealing the influence of knowledge exchange and diffusion. Therefore, it is necessary to compare the results of the three analyses and then make a comprehensive measurement with the integration of the three-dimensional citation network, which is the first major topic of this study.

Despite the controversy over applying simple bibliometric indicators in scientific evaluation, it is undeniable that the citation frequency, the $h$-index, and other factors are still indispensable quantitative indicators when evaluating academic influence. Considering the differences between the evaluation indexes and to avoid the disadvantages of using a single index evaluation, numerous researchers (Moed, 2014; Sidiropoulos, 2016) proposed the idea of comprehensive evaluation. However, research on the comprehensive evaluation of three types of citation network indicators and traditional quantitative indicators is still sporadic, which is the second major topic of this study.

Data and methodology

Basic Data

The field of gene editing emerged in the early 1980s. In the past 10 years, especially since the advent of CRISPR/Cas9 (clustered regulatory interspaced short palindromic repeats/CRISPR-associated protein 9) technology, it has become a popular topic globally. Therefore, the data for this study comprise a set of downloaded records retrieved from Web of Science with “gene edit*” or “CRISPR” as the keywords, covering four types of documents (articles, proceeding papers, reviews, and book reviews). The period covered is from 1980 to 2015. Due to name disambiguation, we manually used authors’ full name and the information of their organization to identify complete author names. Finally, we selected 111 authors whose $h$-index was more than five and papers published amounted to more than 10 as core authors in this field, that is, the research subject of this paper.
Methodologies

Compared with co-citation and bibliographic coupling, direct citation is a direct citing relation without a third-party paper. In this paper, we refer to the weighted direct citations (WDC) proposed by Persson (2010). WDC integrates direct citations strength with weighted co-citations strength and weighted bibliographic coupling strength into a new citation strength. Normalized weighted direct citation (NWDC) is further obtained by normalizing co-citation and bibliographic coupling citation strengths. However, the difference is that the WDC proposed by Persson is demonstrated and applied at the literature level while in this paper it will be applied at the author level. As the author relation is generally based on a specific set of documents, the process is relatively more complex. The principle of the algorithm is depicted in Figure 1, in which the coupling and co-citation frequencies are calculated using the minimum value (Wang, 2014).

![Figure 1. Principle diagram of author weighted direct citation](image)

In Figure 1, $l$ represents the number of times author B is cited by author A; $c$ represents the minimum number of times author A and author B are citing document C (i.e., the coupling frequency of A and B to the citation of C); $m$ represents the total cited frequency of document C; $d$ represents the minimum number of times author A and author B are cited in document D; and $n$ represents the total number of references in document D. In the citation relation depicted in Figure 1, the WDC value of author A and author B is $l + c + d$ and the NWDC value is $l + c/m + d/n$. The Visual Basic for Applications (VBA) program in Excel can process all types of citation analysis data. VBA is a macro language for Visual Basic that Microsoft has developed to perform generic automation (OLE) tasks in its desktop applications.

The $h$-degree of nodes (Zhao, 2011) means that a node has at the most $h$ links in the weighted network, the weight of which is not less than $h$. Hence, it not only considers the number of adjacent nodes but also the strength of the link. Considering this, this paper selects $h$-degree to measure scholars’ comprehensive influence for a single citation network. Due to the difference in characteristics of the core author index, this paper adopts the principal component analysis (PCA) method, the entropy weight method, and Skyline operator (Sidiropoulos, 2016) to measure core scholars’ influence.

The $h$-degree of nodes (Zhao, 2011) means that a node has at the most $h$ links in the weighted network, the weight of which is not less than $h$. Hence, it not only considers the number of adjacent nodes but also the strength of the link. Considering this, this paper selects $h$-degree to measure scholars’ comprehensive influence for a single citation network. However, considering the change of calculation process and network attributes in the integrated network, the paper adopts the centrality indicators to measure the influence of scholars. Based on the idea that the importance of the node not only depends on the number of its neighbor nodes but also the importance of the neighbor nodes, this paper selects in-degree, out-degree, and eigenvector
centrality to measure scholars’ influence in the integrated network. Degree centrality reveals the citing and cited concentration degree intuitively. The eigenvector centrality reflects the author’s own importance by the importance of the associated author. In addition to the above-mentioned citation network indicators, four traditional quantitative indicators (the number of publications, $h$-index, total number of citations, and average cited frequency) are also included in the study, to evaluate scholars’ influence from academic contribution and academic influence two aspects. Due to the difference in characteristics of the core author index, this paper adopts the principal component analysis (PCA) method, the entropy weight method, and Skyline operator (Sidiropoulos, 2016) to measure core scholars’ influence.

Results and discussion

To begin with, the author co-citation matrix, author bibliographic coupling matrix, and author direct citation matrix of core authors in the gene editing field are all established. Next, we adopt the NWDC network (see Figure 2) to measure core authors’ comprehensive influence in three-dimensional citations. The visual analytics examines using Ucinet. To visualize more clearly, we set the threshold to 0.1. As depicted in Figure 2, the overall structure of the NWDC network is divided into two major categories. The lower left network contains lesser nodes but is dense, whereas the upper right has more nodes but is sparse.

![Figure 2. NWDC network](image)

Since the core author direct citation network belongs to the weighted directed network, the NWDC network is also the weighted directed network. Therefore, we choose NWDC network in-degree and out-degree as two indicators to measure core author influence. As shown in Table 1, authors such as Makarova KS, Marraffini LA, and Wiedenheft B are the most cited while those such as Barrangou R, Westra ER, and Makarova KS are the most citing.

<table>
<thead>
<tr>
<th>Num</th>
<th>Author</th>
<th>In-degree</th>
<th>Num</th>
<th>Author</th>
<th>Out-degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Makarova KS</td>
<td>111.557</td>
<td>1</td>
<td>Barrangou R</td>
<td>117.641</td>
</tr>
<tr>
<td>2</td>
<td>Marraffini LA</td>
<td>82.269</td>
<td>2</td>
<td>Westra ER</td>
<td>91.449</td>
</tr>
</tbody>
</table>
Finally, 10 indicators are considered in the comprehensive evaluation system of core authors’ influence. They include the single network measure index (direct citation $h$-degree, co-citation $h$-degree, and coupling $h$-degree), the integrated network measure index (NWDC in-degree, NWDC out-degree, and NWDC eigenvector centrality), and the traditional quantitative index (number of publications, $h$-index, total number of citations, and average cited frequency). However, to some extent, all indicators may reflect overlapping information. It is difficult to objectively evaluate the core authors’ influence directly by using the above indicators. Therefore, the PCA method is used to reduce the dimension of these evaluation indexes, and a few independent indexes (principal components) are obtained. The results of the total variance reveal that the initial eigenvalue of the three components is greater than 1, and the cumulative variance is up to 87.543%. Furthermore, in combination with the gravel plot, the number of principal components is set to 3. In addition, the component load matrix results after orthogonal rotation (as shown in Table 2) show that direct citation $h$-degree, co-citation $h$-degree, coupling $h$-degree, NWDC in-degree, NWDC out-degree, and NWDC eigenvector centrality are closely related to principal component 1; the number of publications and $h$-index are highly correlated with principal component 2; and the total number of citations and average cited frequency are closely related to principal component 3. Simultaneously, the indexes contained in principal component 1 are found to be mainly citation network indexes, which are used to measure scholars’ knowledge exchange and communication in this field; those contained in principal component 2 are mainly paper number indexes, reflecting scholars’ academic contribution; those in principal component 3, are citation quality indexes, reflecting scholars’ academic influence.

Table 2. Component load matrix

<table>
<thead>
<tr>
<th></th>
<th>principal component 1</th>
<th>principal component 2</th>
<th>principal component 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>direct citation $h$-degree</td>
<td>0.959</td>
<td>0.047</td>
<td>-0.012</td>
</tr>
<tr>
<td>co-citation $h$-degree</td>
<td>0.895</td>
<td>0.010</td>
<td>0.147</td>
</tr>
<tr>
<td>coupling $h$-degree</td>
<td>0.828</td>
<td>0.200</td>
<td>-0.122</td>
</tr>
<tr>
<td>NWDC Out-degree</td>
<td>0.890</td>
<td>0.177</td>
<td>-0.072</td>
</tr>
<tr>
<td>NWDC In-degree</td>
<td>0.921</td>
<td>0.033</td>
<td>0.095</td>
</tr>
<tr>
<td>NWDC eigenvector centrality</td>
<td>0.871</td>
<td>0.077</td>
<td>0.107</td>
</tr>
<tr>
<td>number of publications</td>
<td>0.165</td>
<td>0.957</td>
<td>-0.033</td>
</tr>
<tr>
<td>$h$-index</td>
<td>0.066</td>
<td>0.932</td>
<td>0.225</td>
</tr>
<tr>
<td>total number of citations</td>
<td>0.100</td>
<td>0.482</td>
<td>0.852</td>
</tr>
<tr>
<td>average cited frequency</td>
<td>-0.007</td>
<td>-0.069</td>
<td>0.984</td>
</tr>
</tbody>
</table>

According to the score coefficient matrix, we can obtain scores of the three principal components for each core author. After the orthogonal rotation transformation, the three comprehensive indicators are independent of each other and
reveal scholars’ knowledge diffusion and control, academic contribution, and academic influence, respectively. To reveal scholars’ academic influence comprehensively, first use the three principal components as independent indicators of the Skyline operator to identify the Skyline set, that is, 12 outstanding comprehensive influential scholars (see Table 3).

However, ranking is inevitable in scientific evaluation. Although the Skyline operator can select outstanding scholars, it cannot provide their objective ranking. Therefore, to further test the effectiveness of the Skyline operator and obtain the ranking, this paper calculates the weight of each principal component by using the entropy weight method; thus, it obtains each author’s influence score through the standardized data. The top 12 authors are shown in Table 3. In contrast, the outstanding scholars obtained using the Skyline operator and the objective weighting method are found to be broadly consistent, especially for the extremely excellent scholars. It also proves the comprehensiveness and effectiveness of the Skyline operator.

**Table 3. Scholars’ comprehensive influence in the gene editing field**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Author</th>
<th>Sum score</th>
<th>Author</th>
<th>Rank in citation network index</th>
<th>Rank in papers amount index</th>
<th>Rank in citation quality index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Barrangou R</td>
<td>73.81</td>
<td>Barrangou R</td>
<td>2</td>
<td>2</td>
<td>33</td>
</tr>
<tr>
<td>2</td>
<td>Makarova KS</td>
<td>59.12</td>
<td>Brennicke A</td>
<td>82</td>
<td>1</td>
<td>92</td>
</tr>
<tr>
<td>3</td>
<td>Marraffini LA</td>
<td>58.37</td>
<td>Doudna JA</td>
<td>32</td>
<td>3</td>
<td>27</td>
</tr>
<tr>
<td>4</td>
<td>Horvath P</td>
<td>53.33</td>
<td>Gregory PD</td>
<td>111</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>Seeburg PH</td>
<td>52.32</td>
<td>Horvath P</td>
<td>9</td>
<td>40</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>Westra ER</td>
<td>50.55</td>
<td>Makarova KS</td>
<td>1</td>
<td>85</td>
<td>31</td>
</tr>
<tr>
<td>7</td>
<td>Wiedenheft B</td>
<td>50.22</td>
<td>Marraffini LA</td>
<td>5</td>
<td>39</td>
<td>18</td>
</tr>
<tr>
<td>8</td>
<td>van der Oost J</td>
<td>49.57</td>
<td>Miller JC</td>
<td>34</td>
<td>107</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>Doudna JA</td>
<td>47.65</td>
<td>Muramatsu M</td>
<td>36</td>
<td>33</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>Koonin EV</td>
<td>46.34</td>
<td>Nishikura K</td>
<td>23</td>
<td>11</td>
<td>23</td>
</tr>
<tr>
<td>11</td>
<td>Honjo T</td>
<td>44.78</td>
<td>Seeburg PH</td>
<td>50</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>Nishikura K</td>
<td>44.62</td>
<td>Zhang F</td>
<td>77</td>
<td>9</td>
<td>17</td>
</tr>
</tbody>
</table>

**Conclusions**

In this paper, 17 scholars with outstanding influence in the gene editing field are identified using the objective weighting method and the Skyline operator, which corresponds with promoters’ actual situation in gene editing, especially that of CRISPR technology. Meanwhile, some differences are observed between the scholars selected by the two methods. For example, Zhang F et al., without high citation network index scores but having certain influence, cannot be effectively identified using the entropy weight method. Therefore, it is necessary to combine the two methods when selecting outstanding scholars. It is worth noting that more outstanding scholars are identified, such as Makarova KS, Marraffini LA, Horvath P, van der Oost J, and Wiedenheft B, in the introduction of the citation network indexes. Although they are not prominent in the traditional evaluation indexes, they have played an important role in the promotion of knowledge exchange and communication, and their academic contribution and influence should not be ignored. The results also suggest
that citation network indexes are effective supplements to comprehensive influence measurement that use only traditional indicators.

References


Medical Research Breakthroughs and the Public Engagement

Fereshteh Didegah¹ – Qi Wang²

¹fdidegah@sfu.ca
Simon Fraser University, School of Communication, Art and Technology, Vancouver, BC (Canada)
²qiwang@ps.au.dk
Aarhus University, Danish Centre for Studies in Research & Research Policy (CFA), Aarhus (Denmark)

Abstract
This study examines the social web presence of NIH groundbreaking research to discover to what extent the articles have been discussed on the social web, in general, and what type of users and networks are shaped around them on Twitter, in particular. Both altmetric (including tweet counts, Facebook posts, blog and news posts and Wiki citations) and citation counts of articles were examined and a moderate correlation was found between citation counts and altmetric counts. Altmetric indicators correlate more strongly with each other and the highest correlation was found between tweet counts and Facebook post counts. 58% of tweets about the NIH articles are posted by scientists and individual citizens (the general public) and the scientists are the major group tend to tweet these articles for the first time. The general public is mainly interested in tweeting the NIH research on promising medical advances. The retweet network shaped around NIH research was also investigated in order to determine the networking and connections between different type of users and indicate who originate the tweets and who retweet them more; the results show that scientists and citizens are the major groups retweeting the posts originated by other users. Citizens and practitioners were found to have strong networks and they widely retweet each other’s posts.

Conference Topic
Altmetrics

Introduction
Research diffusion mainly has three aspects: research flow from one scientific area to other areas; scientific knowledge dissemination from the academic community to the industry; and research spreading from the academic to the public. Knowledge diffusion could be considered as a signal of innovation where research outcomes applied to industrial production. The diffusion of research from the academic to the public has a more complicated process for different reasons; the general public may not be interested in academic research outputs as they find them too technical to understand; and the research community may not pay much attention to produce or report their outputs in a layman language. However, the public engagement in research outputs has gained increasing attention recently and new methods and tools are proposed to investigate to what extent and how research produced by academia influence the general public. Altmetrics are alternative indicators derived from
the social web platforms that have been proposed to give immediate impact evidence and to help assess the contributions of science to society beyond the scientific community (Priem, Taraborelli, Groth, & Neylon, 2011). Tweet count linking to articles on Twitter is a prominent altmetric indicator showing an immediate impact (Thelwall, Haustein, Larivière, & Sugimoto, 2013). Despite the increasing number of research on altmetrics especially tweet mentions, some questions have yet remained unanswered about the quality and reliability of such platforms in measuring the influence of research outputs on different parts of a society; for example, what type of research is grabbing a higher attention on Twitter? how is research disseminated or discussed on this platform? or who are acting around research articles on Twitter? Although the general public is assumed to be the main active group (Haustein, Tsou, Minik, Brinson, Hayes, Costas, & Sugimoto, 2016), tweets about research may primarily originate from scholars and serve to publicize, rather than discuss, recent research (Thelwall, Tsou, Weingart, Holmberg, & Haustein, 2013).

Scientists are devoting an enormous amount of time and energy to find solutions for problems raised and scientific breakthroughs are increasingly developing to solve the world’s problems. Medical breakthroughs are the most prominent discoveries as they aim at an important aspect of each society i.e. health. The National Institute of Health (NIH) is a primary agency for medical and biomedical research in the United States and the largest public funding body in the world supporting research projects. The agency is investing more than $32 billions a year in the biomedical and health research projects aiming at enhancing life and reducing illness and disability. Research funded by NIH is categorised into three groups: clinical advances in prevention, diagnosis, and treatment of human disease; promising medical advances i.e. potential findings for enhancing human health; and insights from the lab i.e. noteworthy advances in basic research1. NIH research in these categories has led to advancements and breakthroughs in helping people live longer and healthier and they have received top scientific awards like Nobel Prizes; but the question is that are there any societal interactions occurring around these research outputs? To what extent scientific breakthroughs succeed to reach the public attention and solve their problems?

How research is influencing different aspects of society has recently grabbed the attention of research policy makers and funders. Researchers at universities have traditionally been assessed through their research productivity and citation impact but they are or will be also assessed whether their research outputs have any influence on the society and different aspects of it. These are urging scholars to engage in public engagement and social outreach. Altmetric platforms such as Twitter may help to partly answer the questions above through investigating the research disseminated and distributed on Twitter and whether the breakthroughs in science are also important to the general public. Altmetric.com provides information on Twitter user types and categorises users into four groups of ‘member of the public’, ‘practitioner’, ‘science communicator’ and ‘scientist’ but their data has some limitations as through a manual checking, majority of users identified as a member of the public in Altmetric.com are found to be a member of the academia (Tsou, Bowman, Ghazinejad, & Sugimoto, 2015). Few altmetric research addressed this issue and investigated tweeter types through manual checking but their findings vary due to various methods and user classifications used to identify tweeters. A research on 15 top tweeted Finnish articles classified tweeters into 17 categories and reported a high percentage of health care professionals in Medical and Health Sciences, and a high percentage of businessmen? in Social Sciences and Humanities (Vainio & Holmberg, 2017). Another investigation of a random sample of 200 users tweeting research articles classified the users into four groups of ‘brokers’, ‘orators’, ‘broadcasters’ and ‘numblers’ based on two criteria: engagement (tweet content whether it only contains article title or any extra text) and exposure (number of followers). The users were also identified to be an individual or an organisation in each of above

---

categories and the results showed that 68% of the users were individuals (Haustein et al., 2016). But there is no previous study examined the visibility of and interactions around medical research breakthroughs on the social web which is the main focus of current study. This study collects a number of research articles published from the groundbreaking NIH-funded projects and trace them on Twitter to discover to what extent and by whom the articles have been discussed on the platform. The citations to the mentioned articles were collected from the Web of Science (WoS) to compare with tweet counts, Facebook posts, news posts, blog posts, and Wikipedia posts. The main research objective is to investigate to what extent medical research advancements and breakthroughs are discussed on the social web and among the general public acting on Twitter. To reach this goal, two main questions are answered:

1. To what extent is NIH funded research mentioned on different social web platforms versus citation platforms?
2. Which type of users are interacting and engaging around NIH funded research on Twitter? Are they mainly from the public? And who is bringing this research up to Twitter for the first time?
3. Who are the retweeters? Who retweets the tweets posted by scientists or citizens or any other type of users?

Methods
All articles from the three NIH categories including (1) clinical advances in prevention, diagnosis, and treatment of human disease, (2) promising medical advances i.e. findings with potential for enhancing human health, and (3) insights from the lab i.e. noteworthy advances in basic research were retrieved from 2012-2015 accounting for 133 papers, out of which 107 had a DOI. Some articles were classified into two or more categories; so, 37 articles were classified into the first category, 73 articles into the second and 62 articles into the third category. Then, the DOIs were matched with Altmetric.com dataset (last updated version, June 2016) and a number of 16,345 tweets were obtained for all the 107 articles. Author description field of each tweet was extracted from the dataset for further analysis. Altmetric.com provides user types in four groups of ‘member of the public’, ‘practitioner’, ‘science communicator’ and ‘scientist’ but their method to identify user types has some limitations that have been pointed out in previous research; majority of users identified as a member of public by Altmetric.com are actually found to be individual researchers (Tsou, Bowman, Ghazinejad, & Sugimoto, 2015). Moreover, a Twitter account with a blank bio is categorized under ‘public’ rather than ‘unknown’ (Altmetric.com). Hence, six different types of users were designated for this research and each tweet in the dataset was manually checked for its Twitter account to be an/a: individual scientist, individual citizen (a member of the public), organisation, journals/publishers, practitioner, or Blank or other. The results of manual checking for user types was compared with that of Altmetric.com. As seen in Table 1, more than 82% of tweets in each NIH category were found to be a member of the public while individual scientists are the major group tweeting NIH research in all three categories based on the results from manual checking (Table 4). Other altmetric counts including Facebook mentions, blog posts, news posts and Wiki citations were also retrieved from Altmetric.com. The citation counts to articles were obtained from the in-house version of the Web of Science (WoS) provided by Leiden University. The July-2016 version of WoS was used to give both citation and altmetric counts the same time window.
Table 1. User types in different NIH categories according to Altmetric.com

<table>
<thead>
<tr>
<th>User type</th>
<th>NIH category (%tweets)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Members of the public</td>
<td>85.1</td>
</tr>
<tr>
<td>Practitioners</td>
<td>5.4</td>
</tr>
<tr>
<td>Science communicators</td>
<td>3.4</td>
</tr>
<tr>
<td>Scientists</td>
<td>6.1</td>
</tr>
</tbody>
</table>

Results

To what extent is NIH funded research visible on different social web platforms versus citation platforms?

Different social web platforms including Twitter, Facebook, news outlets, blogs, and Wikipedia were examined for mentions of NIH research articles. WoS was also gauged for the signs of scientific impact. As seen in Table 2, NIH research was highly cited (226.2 citations per article on average) and they were also highly tweeted with an average of 178.3 tweets per article. The research had an average of 21.7 news posts, 15.4 Facebook posts and 9 blog posts but they were scarcely cited in Wikipedia. The correlation between citation counts and each of altmetric counts is tested. The results show that number of citations moderately correlate with all altmetric indicators. Number of tweets strongly correlate with the number of Facebook, news and blog posts. Facebook posts are in a strong correlation with news and blog posts and moderately correlate with the number of wiki citations. Number of news posts also strongly correlate with the number of blog posts (Table 3). Across the NIH categories, research that received the highest number of tweets, Facebook posts, news posts and blog posts, were less cited. The results show that research in clinical advances (prevention, diagnosis, and treatment of human disease) had the highest mentions on Twitter, Facebook, news and blogs but they were least cited compared to that of the other two categories. Wikipedia citations positively correlate with WoS citation counts (Table 4).

Table 2. Citation and altmetric impact of total articles

<table>
<thead>
<tr>
<th>Indicators</th>
<th>Tweet</th>
<th>Facebook posts</th>
<th>News posts</th>
<th>Blog posts</th>
<th>Wiki citation</th>
<th>WoS Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average counts</td>
<td>178.3</td>
<td>15.4</td>
<td>21.7</td>
<td>9</td>
<td>0.8</td>
<td>226.2</td>
</tr>
</tbody>
</table>

Table 3. Spearman correlation between citation and altmetric counts

<table>
<thead>
<tr>
<th>Spearman's rho</th>
<th>Tweet</th>
<th>Facebook posts</th>
<th>News posts</th>
<th>Blog posts</th>
<th>Wiki posts</th>
<th>WoS citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweet</td>
<td>1</td>
<td>0.728</td>
<td>0.655</td>
<td>0.663</td>
<td>0.356</td>
<td>0.522</td>
</tr>
<tr>
<td>Facebook posts</td>
<td>0.728</td>
<td>1</td>
<td>0.664</td>
<td>0.625</td>
<td>0.389</td>
<td>0.486</td>
</tr>
<tr>
<td>News posts</td>
<td>0.655</td>
<td>0.664</td>
<td>1</td>
<td>0.679</td>
<td>0.39</td>
<td>0.473</td>
</tr>
<tr>
<td>Blog posts</td>
<td>0.663</td>
<td>0.625</td>
<td>0.679</td>
<td>1</td>
<td>0.436</td>
<td>0.595</td>
</tr>
<tr>
<td>Wiki posts</td>
<td>0.356</td>
<td>0.389</td>
<td>0.39</td>
<td>0.436</td>
<td>1</td>
<td>0.447</td>
</tr>
<tr>
<td>WoS citation</td>
<td>0.522</td>
<td>0.486</td>
<td>0.473</td>
<td>0.595</td>
<td>0.447</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 4. Citation and altmetric impact of articles across NIH categories

<table>
<thead>
<tr>
<th>NIH category</th>
<th>Average post counts of</th>
<th>Tweet</th>
<th>Facebook</th>
<th>News</th>
<th>Blog</th>
<th>Wikipedia</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>226.1</td>
<td>17.5</td>
<td>24</td>
<td>8.9</td>
<td>0.7</td>
<td>167</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>161.2</td>
<td>15.6</td>
<td>20.3</td>
<td>8.8</td>
<td>0.9</td>
<td>263</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>146.5</td>
<td>12.4</td>
<td>17.4</td>
<td>8.5</td>
<td>1</td>
<td>270</td>
</tr>
</tbody>
</table>

Which type of users are interacting and engaging around NIH funded research on Twitter? Are they mainly from the public? And who is bringing this research up to Twitter for the first time?
The 16,345 tweets made to the NIH articles were considered to be manually checked in terms of their users and their results are presented in Tables 5-7. The results showed that in the entire dataset, around 33% of tweets were made by individual scientists. Individual citizens are the second major group posting 25.17% of the tweets and practitioners that are mainly health care professionals and biologists made 13.67% of the tweets to NIH research articles. Journals/publishers tweeted the lowest tweets and had the least contribution (5.87%) (Table 5). Across the three NIH categories, quite the same pattern is observed as around 30% of tweets in category 1, 35% in category 2, and 38% in category 3 are posted by individual scientists. The second major group is also citizens across all three categories (Table 6).
The type of users tweeted research articles for the first time was also examined. The findings showed that more than 40% of articles were tweeted by scientists for the first time. 22% of the articles were also tweeted by journals/publishers for the first time and more than 10% were tweeted first by citizens (Table 7).

Table 5. User types of the total article set

<table>
<thead>
<tr>
<th>User type</th>
<th>% of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual scientist</td>
<td>33.21</td>
</tr>
<tr>
<td>Individual citizen</td>
<td>25.17</td>
</tr>
<tr>
<td>Practitioner</td>
<td>13.67</td>
</tr>
<tr>
<td>Blank or other</td>
<td>12.06</td>
</tr>
<tr>
<td>Organization</td>
<td>10.02</td>
</tr>
<tr>
<td>Journals/Publishers</td>
<td>5.87</td>
</tr>
</tbody>
</table>

Table 6. User types across NIH categories

<table>
<thead>
<tr>
<th>User type</th>
<th>Category 1 (%)</th>
<th>User type</th>
<th>Category 2 (%)</th>
<th>User type</th>
<th>Category 3 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual scientist</td>
<td>28.86</td>
<td>Individual scientist</td>
<td>34.53</td>
<td>Individual scientist</td>
<td>37.54</td>
</tr>
<tr>
<td>Individual citizen</td>
<td>25.39</td>
<td>Individual citizen</td>
<td>26.41</td>
<td>Individual citizen</td>
<td>25.68</td>
</tr>
<tr>
<td>Practitioner</td>
<td>18.07</td>
<td>Blank or other</td>
<td>12.33</td>
<td>Blank or other</td>
<td>13.29</td>
</tr>
<tr>
<td>Blank or other</td>
<td>11.25</td>
<td>Practitioner</td>
<td>10.92</td>
<td>Organization</td>
<td>9.39</td>
</tr>
</tbody>
</table>
Table 7. Percentage of DOIs posted on Twitter for the first time by different user types

<table>
<thead>
<tr>
<th>User type</th>
<th>%DOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual scientist</td>
<td>40.52</td>
</tr>
<tr>
<td>Journals/Publishers</td>
<td>21.55</td>
</tr>
<tr>
<td>Blank or other</td>
<td>11.21</td>
</tr>
<tr>
<td>Individual citizen</td>
<td>10.34</td>
</tr>
<tr>
<td>Organization</td>
<td>9.48</td>
</tr>
<tr>
<td>Practitioner</td>
<td>6.9</td>
</tr>
</tbody>
</table>

Who are the retweeters? Who retweets the tweets posted by scientists or citizens or any other type of users?

The aim is to investigate the network of retweeters for each type of users on Twitter. The results show that 58% of tweets posted by individual citizens are retweeted by citizens. Interestingly, 17% of tweets from citizens are retweeted by scientists and also another 17% by practitioners. Tweets originated by scientists are mainly retweeted by other scientists (45%). 31% of scientists’ tweets are retweeted by citizens. Citizens are the main group retweeting the tweets originated by practitioners (36% of practitioners’ tweets are retweeted by citizens). The tweets originated by journals/publishers are mainly retweeted by scientists (40%) and then citizens (26%). Tweets by organizations are equally retweeted by scientists and citizens (each group retweeted 35% organization’s tweets) (See Figures 1-5). The contribution of citizens in retweeting other tweets is very high which shows that they mainly play their role as a re-distributer and not the main disseminator.

Figure 1. Users who retweet citizens’ tweets

Figure 2. Users who retweet scientists’ tweets
The tweet and retweet networks of the most tweeted article in the dataset

In this section, the most tweeted publication in the dataset functions as an example for conducting an in-depth analysis, which could contribute to understand and examine the role of individual citizens in the discussion of an important medical publication.

The publication tweeted the most is entitled ‘Nurse staffing and education and hospital mortality in nine European countries’, which was published in Lancet, 2014. The paper belongs to the first NIH category, showing that it holds significant promising medical advances. The paper analyzed the relations of nurses’ workloads and the likelihood of patient death and suggesting ‘patient deaths in hospital might be reduced by easing nurses’ workloads and emphasizing education in hiring’. This paper received around 2,200 tweets, which is 12 times more than the number of citations it received. The reason that this paper attracted a high attention on Twitter might be that the paper’s topic is not extremely difficult for non-professionals to understand, and is so relevant to public health. Analyzing users who tweet this paper showed that individual citizens posted most tweets around this paper, accounting for 28% of users, followed by practitioners (26% of total tweets were originated by them). The results show that general public actively participated in the scientific communication around this paper.

For more information about this paper, see: https://www.nih.gov/news-events/nih-research-matters/nurse-staffing-education-affect-patient-safety.
To examine the role of individual citizens in acting around this paper on Twitter, the original tweets and their retweets were investigated. The results show that around 32%, 28% and 19% of tweets originated by scientists, practitioners, and organizations. This result is plausible because both scientists and practitioners are more likely to follow and monitor the development of scientific research. Hence, they may be informed of new and promising publications faster than the public. Individual citizens are the fourth group originating tweets about this paper (Table 8) but they are the major group retweeting the original tweets posted about this article (Table 9), showing that they follow what scientists and practitioners tweet and that they are mainly inclined to retweet their tweets. While scientists, journals/publishers and organizations are the main disseminators of research on Twitter, citizens mainly follow and re-distribute what posted by other users. The top five tweeters who originated tweets about this paper were an organization account, a journal/publisher account and three scientists. The tweets by these five original tweeters were retweeted between 32-84 times. The high retweets could be because of the number of followers that organizations and journals/publishers have that their posts are retweeted more frequently than the other users. Tweets originated by citizens were not retweeted much, showing that other users are not interested in the research shared by this group.

### Table 8. Total number of tweets, original tweets, and retweets by each type of users

<table>
<thead>
<tr>
<th>User type</th>
<th>Total no. tweets</th>
<th>% total tweets</th>
<th>No. original tweets</th>
<th>% original tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practitioner</td>
<td>597</td>
<td>26</td>
<td>164</td>
<td>28</td>
</tr>
<tr>
<td>Journals/Publishers</td>
<td>57</td>
<td>3</td>
<td>25</td>
<td>4</td>
</tr>
<tr>
<td>Organization</td>
<td>242</td>
<td>11</td>
<td>112</td>
<td>19</td>
</tr>
<tr>
<td>Blank or other</td>
<td>223</td>
<td>10</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>Individual scientist</td>
<td>499</td>
<td>22</td>
<td>186</td>
<td>32</td>
</tr>
<tr>
<td>Individual citizen</td>
<td>641</td>
<td>28</td>
<td>86</td>
<td>15</td>
</tr>
</tbody>
</table>

### Table 9. Type of users who retweet the tweets originated about the paper

<table>
<thead>
<tr>
<th>Retweeters</th>
<th>No. retweets</th>
<th>% retweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual citizen</td>
<td>518</td>
<td>32.85</td>
</tr>
<tr>
<td>Practitioner</td>
<td>405</td>
<td>25.69</td>
</tr>
<tr>
<td>Individual scientist</td>
<td>300</td>
<td>19.02</td>
</tr>
<tr>
<td>Blank or other</td>
<td>201</td>
<td>12.75</td>
</tr>
<tr>
<td>Organization</td>
<td>121</td>
<td>7.67</td>
</tr>
<tr>
<td>Journals/Publishers</td>
<td>32</td>
<td>2.03</td>
</tr>
</tbody>
</table>
Discussions and conclusion
Twitter seems a primary social web to spread academic research and the NIH groundbreaking articles are highly mentioned on this platform compared to the other social web platforms. The articles even received more citations than tweet counts and while they were cited 226.2 times on average, they were tweeted 178.3 times on average. Twitter intensity of NIH articles is 5 times higher than that of reported in Didegah, Mejlggaard, and Sorensen (in press) for a sample of random articles in biomedical and health research (34.3 tweets per article). The NIH research received 15.4 mentions on Facebook, 21.7 mentions in blogs, and 9 mentions in news on average but they have been very rarely cited in Wikipedia. Previous research also shows higher counts of tweets to research articles than the other social web mentions (Thelwall, Haustein, Larivière, & Sugimoto, 2013). A moderate correlation was found between number of citations and altmetric counts. However, altmetric indicators highly correlate with each other and the highest correlation was found between number of tweets and number of Facebook posts. The citation and altmetric counts to articles vary across NIH categories. The articles representing clinical advances in prevention, diagnosis, and treatment of human disease are more mentioned on the social web and the articles representing insights from the lab are mentioned the least. This is vice versa for citations, as the articles disseminating the insights from the lab are more cited than the articles representing clinical advances and also promising medical advances. This may show that the social web users are more interested in the clinical advances in prevention, diagnosis, and treatment of human disease, while scientific community tends to cite the lab insights. The type of users tweeting the articles showed that well above 58% of tweets of the articles were posted by scientists and the lay people, respectively. Practitioners contributed to 13.7% of tweets posted about the articles. Two recent research showed that medical research is majorly tweeted by practitioners (Vainio & Holmberg, 2017; Didegah, Mejlggaard, & Sorensen, in press). The difference may be due to different samples used; while groundbreaking articles among the medical research community are the focus of this study, both previous studies examined samples of highly tweeted medical articles. Journals and publishers are the minor group tweeting articles (they only did 6% of tweets). Scientists are the major group tweeting articles from all three NIH categories also. The high average of citations per article may be a consequence of high activities around the articles on Twitter by the scientists. Interestingly, the second major group tweeting articles on promising medical advances is individual citizens while the second major group tweeting articles on clinical advances and lab insights are practitioners, showing a public interest in the articles on the promising medical advances. The individual scientists are also the first group tweeting the NIH groundbreaking articles. Some of these scientists may probably be the authors of the articles publicizing their research on Twitter right after publication. About 22% of tweets were posted by journals/publishers for the first time. Around 56% of the NIH groundbreaking articles are published in Nature and Science, respectively. These two journals have active Twitter accounts and tweet their recent articles frequently. This may show the role of journals activities on the social web in informing their recently published articles online.

The retweet network shows that scientists and citizens are the major groups retweeting the tweets originated by the other users. Citizens are the major group retweeting the posts originated by practitioners and practitioners are also the main group retweeting the tweets by citizens which shows the strong connection between these two groups on Twitter in sharing research publications. Furthermore, the citizens who are interested in practitioners’ tweets and take an action and retweet their posts could be mainly among patients. This shows that practitioners and patients are widely following each other and have strong networking on Twitter.

The high engagement of citizens in acting around NIH research breakthroughs may be because first, these papers are written in a language that is understandable for lay people and second, the topics and problems discussed are of a public interest.
In short, medical breakthroughs identified through NIH groundbreaking funded research are widely mentioned on Twitter. The articles are also highly cited; an external factor that positively contributed to high citations of these articles could be that the articles are highly tweeted by scientists and also that these articles are mentioned on Twitter for the first time by the scientists. The lay people also showed a wide interest in acting around the NIH research on Twitter and they are mainly interested in the research on the promising medical advances. In the next phase of this study, the tweets about the NIH research will be further investigated through a content analysis in order to identify what the users are saying about these articles and how deeply they are engaging in the articles content and results.

References
Neologisms and affiliation name: a bibliometric perspective on the institutionalization of interventional radiology.

Philippe Gorry

philippe.gorry@u-bordeaux.fr
GREThA UMR CNRS 5113, University of Bordeaux, (France)

Abstract
Delimitation of scientific subfields constitutes one of the key problems in scientometrics. De Bruin and Moed (1993) suggested that “corporate address words” or word co-occurrence in affiliation field could be used. In science, researchers construct knowledge through neologisms. Studies of new scientific concepts indicated us new trends on naming institutions with neologisms. Therefore, one can postulate that neologisms are markers of emerging scientific field, and naming institution after neologisms could be part of a legitimation process. The aim of this paper is to explore this hypothesis through the bibliometric analysis of one case study related to the diffusion process of the concept of "interventional radiology" (IR). The study of the history of IR concept and its trends using affiliation addresses allows us to better understand how IR field has attempted to institutionalize and to gain its autonomy. Several institutions publishing in the field of IR have undertaken a legitimization process by entitling department "interventional radiology", but the specialty has partially or internationally failed to gain its scientific autonomy. In conclusion, measuring the trend of newly named institutions after a neologism can be used to assess the degree of institutionalization for a new medical or scientific field.

Conference Topic
Science of science

Introduction
Affiliation name, neologism and institutionalization
Delimitation of scientific subfields constitutes one of the key problems in bibliometrics. Several methods have been proposed based on indexing systems, co-citation and co-word analysis. De Bruin and Moed suggested in 1993 that “corporate address words” are words referring to scientific (sub)fields and then, can be used to delimitate research field. However, since his pioneering work on corporate addresses in publications with the study of words co-occurrence in the affiliation field, his idea didn’t receive much attention in the bibliometric literature with the exception of the literature on the affiliation disambiguation problem and institutions ranking. The limitation of this approach is that the affiliation addresses don’t contain cognitive words with high frequency, beside words such as university, faculty, mathematics or biology for examples. However, science and technology are producing a plethora of new words or neologisms (Fernandez-Silva, 2016). A neologism is a word term, or phrase which has been recently coined to apply to new concept or to make older terminology more contemporary. In science, researchers construct knowledge through term formation, and used them as rhetorical device in scientific papers, especially at a time when scientific understanding is exploding, producing entirely new scientific field (Pecman, 2014).
Preliminary results of ongoing scientometric analysis on the study of the career of concepts such has “health technology assessment” (Benoit & Gorry, 2017) or “regenerative medicine” (Kawecki & Gorry, 2015) indicated us new trends on naming institution with neologism related to these emerging medical or scientific fields. Therefore, one can postulate that neologisms are markers of emerging scientific field, and naming institution after neologisms could be part of a legitimation process of new research field. Legitimation or legitimacy refer in social sciences to the process of making something normative in the society. It's a public search for a social identity becomes widespread in a group whenever the status of new scientific field has yet to win acceptance or is under attack (Merton, 1973).
The aim of this paper is to explore this hypothesis through the bibliometric study of the capture of one neologism in affiliation address. This new bibliometric indicator could be seen as a measure of the degree of institutionalization (Raina & Gupta, 1988).

Case study
Interventional radiology (IR) was born in the 1960s, but its institutional recognition was late and incomplete. IR is still the subject of turf wars between different medical disciplines and it is not clear if this field will one day become an independent medical specialty (Rosch, 2003). The aim of this study is to analyze the diffusion process of the concept of IR using a scientometric approach. The study of the history of this concept and its trends allows us to better understand how this medical field has attempted to institutionalize and to gain its autonomy. The term “interventional radiology” is widely used by many individuals, academic institutions, professional bodies, and funding organizations. It has become an accepted neologism despite the lack of an agreed upon clear or precise definition (Mignot & Gorry, unpublished results).

Method
A literature search on “interventional radiology” concept was conducted in PubMed database. First, a search query was conducted with the medical subject heading (MeSH) “radiology, interventional” as well complementary terms such as “interventional magnetic resonance imaging”, “interventional radiography”, and “interventional ultrasonography”. But the MeSH was not sensitive enough: while the term of “interventional radiology” was first connoted by Wallace in 1976, it was only referenced in the thesaurus in 1990. Therefore, all the keyword strings were associated with a disjunction boolean operator, and searched within the title or abstracts fields. In order to control the validity of the obtained corpus, we tested the occurrence of the concepts in the documents. Unsurprisingly, “coronary angiography”, “angioplasty”, “catheter”, that define IR field very well, were the most recurrent terms. Therefore, the publications corpus was closely related to the “Interventional radiology” field. A corpus of 26655 documents in all types of records were harvested through 31 December 2014. Second, a search query was conducted within the affiliation field with the term “interventional radiology”. A second corpus of 6616 publications were harvested through 31 December 2014. Both corpus was extracted from Pubmed® database and imported, parsed and disambiguated for author and affiliation names with the help of Intellixir®, a statistical content analysis and text mining software-as service (SaaS) in order to build an in-house database (Gorry & Ragouet, 2016). The disambiguation, is based on a lexical similarity approach with a three-gram based splitting technique. During a first step, the SaaS automatically disambiguate institution names in affiliation field with a similarity score higher than 90%; at the end, a post-processing step allow the investigator to manually aggregate or not the remaining affiliation data. Then, we analysed the resulting corpus by conducting descriptive statistics analysis and drawing choropleth map using Intellixir® (Benoit & Gorry, 2017). Thus, we were able to examine the history of the concept, identify the actors involved in IR (authors, journals, institutions and countries) and understand the institutionalization of this medical field.

Result
Analysis of “interventional radiology” concept through keywords, abstract and MeSH
The data obtained were first organized to build a timescale, corresponding to the growth of publications referring to “interventional radiology” for the period retained (Fig. 1). While interventional radiology is born with the invention of angioplasty by Dotter in 1964 (Gorry & Ragouet, 2016), Pubmed® recognize a first publication titled “Interventional radiology”
published by Wallace in 1976. The temporal distribution of publications suggests a peripheral status of interventional radiology in the literature until a turning point in the late 1980’s. From that time, the growth of the publications has become exponential, and reached more than 2000 articles in 2010. Four general steps can be dissociated on the curve: 1976-1987, 1987-1992, 1992-2001, & 2001-2014. It can be seen that the occurrences of the concept in the literature did not significantly progressed until 1992. Then, after a period of growth, the use of the concept step down. Since 2001, the term use at a faster growth rate reflecting a new step.

**Figure 1. Trend of publications referring to “interventional radiology”**.

![Graph showing the trend of publications referring to “interventional radiology”](image)

**Analysis of “interventional radiology” concept through affiliation field**

Among the corpus of documents harvested through the use of “interventional radiology” in the affiliation field, we identified 100 different institutions after disambiguation. During the same period, the number of new institutions publishing in the IR field, regardless of their denomination, was 754. The first publishing institution with affiliation entitled with “interventional radiology” term appeared in 1987: it was the Department of Radiology, Angiography and Interventional Radiology at Brigham and Women's Hospital (US).

**Figure 2. Trends of new institution entitled “interventional radiology”**.

![Graph showing the trends of new institution entitled “interventional radiology”](image)
Then, the number of new institution entitled “interventional radiology” year after year was measured (Fig. 2). The number of new IR named institution reached a maximum (n=10) in 2001 and followed a rapid decline with only one new IR named institution in 2013 (Fig. 2: black box). In parallel, one can visualize through the measure of the cumulated IR name institution number that this phenomenon is reaching a plateau (Fig. 2: dashed line). These trends parallel, with a time lag, that of new institutions publishing in the IR field with a maximum (n=71) in 1994, and a rapid decline thereafter (data not show).

Figure 3. World distribution of institutions entitled “interventional radiology” by country.

The contribution of different countries to this trend to name institution after “interventional radiology” was estimated by the location of the IR affiliated institutions with at least one publication in the field of IR, and the top countries were ranked by percentage of IR publications over the total IR publications number (n=6616). As shown in the pie chart (Fig. 3), USA rank first with 34% of IR publications, Germany published the second-highest number of total publications with 15%, followed by China, Italy, and Japan. Out of the 20 top countries, two were emerging economies, China and Turkey. This distribution should be compare to the one for IR publication field as a whole whether the institutions are titled or not with IR (Mignot, L. & Gorry, P. unpublished results).

Figure 4. World map of institution entitled “interventional radiology”.
A cartographic analysis of spatial distribution of “interventional radiology” named institutions allowed us to get a better visualization of the geographic distribution based on the zip code information, independent of the ratio of IR publication by institution (Fig. 4). The comparison of geographical trends concerning the use of “interventional radiology” term in affiliation confirms the important number of IR named institutions in the top countries, but also reveals the implantation of such IR named institutions in many countries around the world except for Africa, including developing countries such as Brazil, India, Russia, …

Table 1. Main institution entitled “interventional radiology” by IR publications number.

<table>
<thead>
<tr>
<th>Rank IR named</th>
<th>Rank IR publication</th>
<th>Name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>Duisburg-Essen University: Dpt. of Diagnostic &amp; Interventional Radiology &amp; Neuroradiology Univ. Hospital Essen</td>
<td>DE</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>Frankfurt University: Dpt. of Diagnostic &amp; Interventional Radiology Clinic of Johann Wolfgang Goethe</td>
<td>DE</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>University of Chicago: Dpt. of Radiology Section of Interventional Radiology.</td>
<td>US</td>
</tr>
<tr>
<td>6</td>
<td>42</td>
<td>Sha Tin University: Dpt of Imaging and Interventional Radiology Prince of Wales Hospital.</td>
<td>HK</td>
</tr>
<tr>
<td>8</td>
<td>22</td>
<td>Guangzhou University: Dpt. of Interventional Radiology Nanfang Hospital Southern Medical University</td>
<td>CN</td>
</tr>
<tr>
<td>12</td>
<td>43</td>
<td>Pisa University: Dpt. of Diagnostic &amp; Interventional Radiology</td>
<td>IT</td>
</tr>
<tr>
<td>13</td>
<td>11</td>
<td>Vienna University: Dpt. of Radiology of Division of Cardiovascular and Interventional Radiology</td>
<td>AT</td>
</tr>
<tr>
<td>20</td>
<td>90</td>
<td>Zagreb University: Interventional Hospital</td>
<td>HR</td>
</tr>
</tbody>
</table>

Finally, we looked at the scientific productivity of the institution named after “interventional radiology” by measuring the number of publications referring to IR term as keyword (Table 1). The ranking of the top IR named institution shows that the most productive institutions were Duisburg-Essen University, followed by the Frankfurt University, and then by the University of Chicago. Then, the first IR named institution by country is reported. We identified among the top 20 IR named institution, department of radiology belonging to university and/or hospital in the main countries (Germany, USA, China) but also in Honk Kong, Austria or Croatia. However, the ranking of these IR named institutions should be relativized by looking at their ranking in the whole corpus of IR publications (Table 1: column Rank IR publication). Some discrepancy could be observed especially in the case of the Sha Tin University or the Zagreb University. Indeed, top institutions publishing IR related articles are not titled after IR term.

Discussion

The promotion of the concept of "interventional radiology" and the attempt of institutionalization of a disciplinary identity has taken place several years after the first medical developments of the specialty. This is confirmed by the dynamic of scientific publications. However, its institutional recognition was late and incomplete. To this day, interventional radiology remains a heterogeneous field with unstable borders. All of these might explain why several institutions have undertaken a legitimization process by entitling department "interventional radiology". The observed trend varies from country to country across advanced and developing economies, the (re-)named institution after IR are struggling to reach the top publishing institutions, and this phenomenon is declining on the overall. This could be interpreted through the theory of innovation diffusion, when a new technology or a market share
is reaching a saturation level (Rogers, 1965). This is due to its hybrid origins (radiology, cardiology, surgery) (Gorry & Ragouet, 2016), but also to the fact that interventional radiologists are still divided regarding choices to make for their future (to remain part of radiology or to emancipate as an independent specialty) (Rosch, 2003). Institution (re-)naming through neologisms might be considered as a branding strategy used by scientific community and academic institution competing for identity and resources (Drori et al, 2013). It refers to theories of organizational identity and institutionalism. Our results highlight the role of social strategies in constructing scientific knowledge and nourish the debate between philosophers and sociologists of science (Fueller, 2000; Longino, 2001).

In conclusion, measuring the trend of newly named institutions after a neologism can be used to assess the institutionalization of new medical or scientific field and to test hypothesis based on qualitative approaches (historical, sociological or political) or quantitative models integrating economic indicators (Benoit & Gorry, 2013). It would open new avenues to explore kuhnian scientific revolution in modern science and nourish history and sociology of science.

Acknowledgments

This work is supported by France’s Institut National du Cancer as part of Project #6165 titled ‘‘Interventional radiology in oncology: a multidisciplinary approach to medical innovation and its social recognition’’, and project #2015-179 titled “Comparative analysis of three translational processes: radiofrequency, HIFU and magnetic hyperthermia cases”. The author would like to thank Emilie Bisbau for her assistance.

References


Scientometrics for informing priority setting in health research: the case of mental and behavioural disorders

Alfredo Yegros,¹ Maria Francisca Abad,² Paula Adam³ and Ismael Ràfols⁴

¹a.yegros@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (The Netherlands)

²abad@uv.es
Dept. of History of Science and Documentation, Faculty of Medicine, Universitat de València, València (Spain)

³padam@gencat.cat
Catalan Agency for Health Information, Assessment and Quality (AQuAS), Barcelona, Catalonia (Spain)

⁴i.rafols@ingenio.upv.es
Ingenio (CSIC-UPV), Universitat Politècnica de València, València (Spain)
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (The Netherlands)

Abstract
In this study, we analyse the publication portfolio on mental and behavioural disorders to explore how this information can inform priority setting in grant allocation for this disease area by the Department of Health of the Catalan government. We compare the disease burden, publications and funding for each mental health condition, using data from the Global Burden of Disease (WHO), Web of Science (WoS) articles selected with MeSH terms (linking WoS with PubMed) and WoS funding acknowledgement. The data obtained suggest that Catalonia underinvests in unipolar depression in comparison to bipolar disorder and schizophrenia. It also highlights that these areas of mental health research are mainly focused on psychiatry and substantially funded by pharmaceutical companies (about 20%-40% of publications).

Conference Topic
Science policy and research assessment

Introduction
Public policies in science are partly shifting their focus towards supporting research that helps address societal problems or grand challenges. According to various analyses, such as Sarewitz (1996), Chalmers et al. (2014) or Stirling (2015), there is a significant misalignment between the focus of much mission-oriented research and the societal needs or demands that such research purports to address. Yet, while many analysts may agree with the idea of better aligning science supply with societal needs (Sarewitz and Pielke, 2007), the operationalisation of such comparison remains challenging.

On the science supply side, one can produce estimates of research portfolio contents on the basis of resources (people, funding) invested in research areas or in terms of research outputs such as publications (Wallace and Rafols, 2015), but it is not obvious how to classify science inputs or outputs into categories that relate to social demands. On the side of societal demands, making estimating social needs is even more difficult and controversial, as it is highly dependent on problem framings (e.g. see Lawrence (2004) on obesity).

Health research is unique in having some data available to estimate both supply and demand. Health related articles have been tagged with medical descriptors such as MeSH (PubMed’s Medical Subject Headings) that relate publications to health conditions. And Global Burden of Disease is an estimate of a societal need for a given health (Murray and López, 2013), which has been institutionalised by the World Health Organisation (WHO) and thus offers a
platform for comparison. Using these data, Agarwal and Searls (2009) and Evans et al. (2014) compared the disease burden for disease groups against the number of publications (matched by MeSH), at the global level. The clearly showed that global publication patterns mainly respond to the health needs of the developed countries. Similar studies by Lewison and colleagues had compared publications with disease burden for specific diseases such as malaria (Lewison and Srivastava, 2008) or cancer in India (Lewison and Roe, 2012) using title keywords to capture relevant papers.

In order to support priority setting in health research, besides providing information on existing publications focus and on disease burden, current degree of funding for certain is also relevant. To analyse degree of funding on certain diseases, Grassano et al. (2017) looked at funding acknowledgement in cancer research in the UK (see also Paul-Hus et al., 2016).

In this study, we analyse the publication portfolio on mental and behavioural disorders to explore how this information can inform priority setting in grant allocation for this disease area by the Dept. of Health of the Catalan government. By examining the distribution of disease burden, publications and funding across ten different mental disorders, we provide decision makers with information that helps them allocate grant with public health considerations.

**Methods**

We obtained the global burden of disease estimates from a dedicated WHO website, using 2012 data.¹ This data source provides estimates of burden in terms of *Disability-Adjusted Life Years* (DALYs), which sums the years not lived until life expectancy age due to deaths produced by a condition (*Years of Life Lost*) and the time not fully enjoyed because of a health condition (*Years Lost due to a Disability*). These estimates are given in categories defined in terms of the Internal Classification of Disease 10 (ICD10)².

We matched the ICD10 categories with MeSH terms using various sources. As main source to establish the correspondence between ICD10 and the Medical Subject Headings (MeSH) we used the Unified Medical Language System (UMLS), produced by the U.S. National Library of Medicine. UMLS does not always provide the equivalence between the two classification systems and therefore in some cases we had to find the closest MeSH for a given ICD10 code. The ICD10 category of ‘Endocrine, blood and immune disorders’ was not considered in this study due to unclear correspondence between ICD10 and MeSH descriptors. In many cases, there is a one-to-one correspondence, but in a few cases one ICD10 category needs to be covered by various MeSH terms. Following Agarwal and Searls (2009), we removed publications from categories to avoid counting twice a publication in two diseases. In this case, publications were left in the MeSH term closer to root of the MeSH tree (e.g. cancer related publications were removed from other disease categories and kept in ‘Malignant neoplasms’).

Publications were obtained from CWTS inhouse version of the Web of Science (WoS) for the period 2009-2013 (5 years) on the basis of their MeSH terms. To do so, we relied in an existing matching in WoS between WoS and MEDLINE/PubMed records, which allows us the use of all the fields in both databases for the same publication.

² [http://www.icd10data.com](http://www.icd10data.com)
Results
The management of a research portfolio should consider a variety of issues, including the diversity of topics (here health conditions) covered, its alignment with perceived societal needs, the possibilities for synergist interactions among topics and the perceived scientific opportunities of a topic at a given moment (Wallace and Rafols, 2015). The analysis below focuses on the alignment as when comparing disease burden and publications, or the diversity of disciplines, but let us stress those broader issues should be considered (which we hope to discuss in the final form of this preliminary paper).

We first compare the disease burden and publications across broad fields, in order to estimate the relative weight of mental health research in Catalonia, benchmarking it against the world, the EU-15 (mainly the western and southern countries of the European Union) and Spain, as shown in Figure 1. Since there are no estimates of disease burden for Catalonia, its publications can be compared to EU-15 and Spain’s disease burden. Regarding mental health, one can observe that Catalonia has a proportion of publications which is substantially lower than its disease burden, but a bit higher than the EU-15 or Spain. Hence, Catalonia is a bit specialised in mental health compared to its European neighbours.

Figure 2 shows the relative disease burden and relative publication focus for the area of mental and behavioural disorders. Childhood behavioural disorders and mental disability are estimated to have a low disease burden in relation to the publications devoted to these conditions. Drug use disorder and schizophrenia are areas in which Catalonia has relatively more publications than burden, with a pattern similar to the world, whereas in unipolar disorder in an area of potential underinvestment. Bipolar disorder is a condition that has relatively many publications in Catalonia in relation to disease burden – a situation that is not observed across the world or for EU-15. Other conditions do not show a strong difference between relative publication and relative disease burden.

In terms of disciplines, ‘Psychiatry’ is found in Figure 3 the dominant Web of Science Category in research on mental disorders in Catalonia. Further research is needed to see if this is also the case in other European countries. Other relevant disciplines are clearly related to specific health conditions, for example, ‘Substance Abuse’ for drug and alcohol use disorders, or ‘Genetics and Heredity’ for idiopathic intellectual disability. Most research is carried out by universities, in particular for schizophrenia, bipolar disorder and unipolar depression (Figure 4). However, hospitals also carry out or collaborate in 40% or more of the publications all other conditions.

Finally, the Spanish Health Funding Agency (Instituto de Salud Carlos III) is found to be main funder, as shown in Figure 5. Funding from the Catalan government is slightly higher (in number of publications including acknowledgement) to that from other Ministries of the Spanish government. About 20% of publications acknowledge funding by pharmaceuticals – 45% in the case of bipolar disorders.

Discussion
This study has aimed at providing information that supports priority-setting regarding research on mental disorders in Catalonia. We have provided various perspectives about current research focus: the relative burden of mental health as compared to publications, the

3 We included under the Health Funding Agency also other funding acknowledged to the Ministry of Health, who manages the agency, given ambiguities in acknowledgements.
relative disease burden of each mental or behavioural disorder as compared to publications, the disciplinary focus of mental health research, the main research performers and the main funders. To our knowledge this is the first study combining disease burden, number of publication and funding acknowledgements.

The main insights gained are the research that Catalonia underinvests in unipolar depression in comparison to bipolar disorder and schizophrenia. We also found that mental health research is mainly focused on psychiatry and substantially funded by pharmaceutical companies (about 20% of publications). In the ISSI conference, we plan to present views by different stakeholders on what this information suggests in terms of desired future research priorities by the Catalan government in mental health.

References


Chalmers, I., Bracken, M. B., Djulbegovic, B., Garattini, S., Grant, J., Gülmezoglu, A. M., ... & Oliver, S. (2014). How to increase value and reduce waste when research priorities are set. The Lancet, 383(9912), 156-165.


Figure 1. Proportion of disease burden (in red, left) vs. proportion of publications (blue, right) on disease groups, for the world, EU-15, Spain and Catalonia.

Figure 2. Proportion of disease burden (in red, left) vs. proportion of publications (blue, right) on various “Mental and behavioural disorder”, for the world, EU-15, Spain and Catalonia.
Figure 3. Number of publications on mental and behavioural disorders in Catalonia (2009-2013) by subdisciplinary category (according to Web of Science Category classification).

Figure 4. Number of publications on mental and behavioural disorders in Catalonia (2009-2013) by main universities (blue, left) and hospitals (red, right).
Figure 5. Proportion of publications with funding acknowledgement on mental and behavioural disorders in Catalonia (2009-2013) by funding source.
Cross-boundary collaboration of highly-cited scientists:

A bibliometric study on scientific publication in agricultural science

Yong Zhao\(^1\) Dong Li\(^1\) Yishan Wu\(^2\) Mingjie Han\(^1\) Chenying Li\(^1\)

\(^1\)zhaoyong@cau.edu.cn
Library of China Agricultural University (China)
\(^2\)wuyishan@istic.ac.cn
Chinese Academy of Science and technology for Development (China)

Abstract
In the recent decades, the growing importance of interdisciplinary fields, such as biotechnology and new materials, has been witnessed constantly. Such researches often involve the collaboration of not only transcending national boundaries or disciplinary boundaries, but also transcending sectors, such as universities and industry. Highly-cited scientists represent some of world’s most influential scientific minds, so analyzing the characteristics of their research activities might provide insights about preferential conditions that foster high-impact work. Regarding the shortage of the researches on the characteristics of scientific collaborations of highly-cited scientists, this study is hoped to contribute some information in this understudied subject by examining the collaboration characteristics of high-cited agricultural scientists.

Key words
Highly-cited Scientists· Cross-boundary collaboration· Agricultural Science· International collaboration· Interdisciplinary collaboration· University-industry-government collaboration

Conference Topic
Studies on the level of individual scientists

Introduction
Scientific collaborations play a vital role in research and development (R&D) processes by communication and knowledge-sharing practices (Clarke 1967; Qin 1997; Beaver 2001; Leimu and Koricheva 2005; Wagner & Leydesdorff 2005; Wuchty et al., 2007; Leydesdorff and Wagner 2008; McKelvey et al. 2015). The previous findings show that collaboration procedures vary according to different needs, such as specialized knowledge or skills, access to high-end equipment, and data or materials (Bozeman and Corley 2004; Melin 2000; Thorsteinsdottir 2000).

In recent decades, the growing importance of interdisciplinary fields, such as biotechnology and new materials, has been witnessed constantly. Such researches often involve the collaboration of not only transcending national boundaries (Yong et al. 2016) or disciplinary boundaries (Peng and Haoxiang 2015), but also transcending sectors, for example, universities and industry (Hema and Raiah 2016). In addition, there are various political factors encouraging greater level of collaboration among
researchers (Katz and Martin 1997). National Science Foundation (NSF) initiated the industry/University Cooperative Research Centers (U/UCRC) program to develop long-term partnerships among industry, academe, and government. And almost all research universities in the USA and Europe have established technology transfer offices that connect the university and private sector (Siegel et al. 2007). Chinese government also issued a series of plans on national scientific and technological innovation, explicitly putting forward to boost university to participate in the collaborations with public research institutes and enterprises (Yin et al. 2016). All of these policies draw on the assumption that strong connections between universities and the private sector are necessary for innovation and economic development (Etzkowitz and Leydesdorff 1999).

Highly-cited scientists represent some of world’s most influential scientific minds, then understanding the characteristics of their research activities might provide insights about preferential conditions that foster high-impact work. Meanwhile, the achievements of eminent scientists offer exemplars and benchmarks by which the rest of us can gauge our own career trajectories and successes (Hermanowicz 1998, 2005). Prior studies have focused primarily on their academic age, career change, lifestyle choices and work habits (John et al. 2010; Deng 2014; Miao et al. 2013). However, very little is known about the characteristics of scientific collaborations of highly-cited scientists. This study is hoped to add some information to this understudied subject by examining the collaboration characteristics of high-cited agricultural scientists.

Concepts and Framework
The rhetoric of boundaries is signified by spatial metaphors of turf, territory, and domain (Klein, 1996). Metaphors of place call attention to the ways categories and classifications stake out differences. In studies of science, where the concept of boundary work arose, the primary focus has been disciplinary formations (Gieryn 1983). In studying the Social Science Research Council, Donald Fisher applied the concept to interdisciplinary activities (Donald 1990). Boundary-crossing behavior occurs in all knowledge fields though, because the problem of boundaries is universal. Boundaries are generally based on knowledge, technology, geography, or some combination of these attributes (Ramlogan et al, 2007; Carlsson, 2006; Malerba, 2005; Wolfe and Gertler, 1998; Cooke, 1997). Some scholars discuss research collaboration as transcending geographic boundaries, motivated instead by a particular problem (Ramlogan et al, 2007; Carlsson, 2006; Malerba, 2005). But multidimensional concept of proximity goes beyond mere geographic closeness to include individuals who are institutionally proximate by sharing organizational structures, regulations, or cultures (Lander, 2013; Holden, 2008). Indeed, Cross-boundary collaboration attempts to break down those boundaries existing in geographic space, scientific disciplines and industrial organization for knowledge sharing.

Cross-boundary collaboration is depicted in Figure 1 as three dimensions (ranging from narrow to broad scope) on which scientific collaboration, regional coalitions, and inter-sector partnerships can be compared. The scope of disciplinary dimension,
incorporated in Figure 1, encompasses three key concepts related to the boundaries of disciplines, multidisciplinarity, interdisciplinarity and transdisciplinarity (Stokols et al, 2003; Wagner et al, 2011). For both the organizational and geographic dimensions, the narrower levels are nested within each of successively broader levels of collaboration (Stokols, 2006). In this article, we combine the framework with bibliometric methods to discover the characteristics of cross-boundary collaboration of highly-cited agricultural scientists.

**Methods and Data**

**Methods**

**International Collaborations**

International collaboration in science can be considered as a communications network that is different from national systems and has its own internal dynamics (Gibbons et al. 1994; Ziman 1994). Based on the thought of fractional count, Wang et al. introduced the International Collaboration Activity Index (ICAI), including five sub-indicators, Average Number of Collaboration Countries per Paper, Paper Collaboration Ratio, Collaborative Leadership Index, International Collaboration Range, and Publishing Paper Ratio, to comprehensively measure the degree of international collaboration at the country or region level (Wang et al. 2014). In this article, we try to apply ICAI indicator to measure the international collaboration of highly-cited scientists at individual level.

\[
\text{ICAI} = \text{ANCCP} \times \text{PCR} \times \text{CLI} \times \text{ICR} \times \text{PPR}
\]

\[
\text{ANCCP} = \frac{\sum_{h=1}^{n} P_n}{\sum_{h=1}^{N} P_n}
\]

\[
\text{PCR} = \frac{\sum_{n=2}^{N} P_n}{\sum_{n=1}^{N} P_n}
\]

\[
\text{CLI} = \frac{\sum_{n=2}^{N} P_n}{\sum_{n=2}^{N} P_n}
\]

\[
\text{ICR} = \frac{TN}{PP}
\]

In the formula:

Average Number of Collaboration Countries per Paper (ANCP) is defined as the number of times the highly-cited researchers have cooperated with authors from the other countries or regions, and reflects the breadth of one highly-cited researcher in
international collaboration from the microscopic view. Among them n represents the number of countries or regions in a paper. (n-1) is regarded as the number of times of international collaboration. And Assuming that a highly-cited researcher P owns Pn papers completed by authors from n countries or regions and N is the maximum value of n (N≥n). \( P_n^{RP} \) represents the number of papers of highly-cited researchers as corresponding authors.

Paper Collaboration Ratio (PCR) is defined as how much a highly-cited researcher’s multinational papers accounted for his or her total number of papers.

Collaborative Leadership Index (CLI) reflects how a highly-cited researcher is motivated with regards to international collaborations. \( \sum_{n=2}^{N} P_n^{RP} \) is the total number of multinational papers that highly-cited researcher P has taken part in as the corresponding author. \( \sum_{n=2}^{N} P_n \) represents the number of multinational papers a highly-cited researcher P has taken parts in.

International Collaboration Range (ICR) is defined as how many partner countries or regions have been involved in collaborations, which can be described as an absolute value. \( TN \) is the total number of countries or regions with which a highly-cited researcher P has cooperated. This indicator reflects the breadth of one highly-cited researcher in international collaboration from the macro view.

Publishing Paper Ratio (PPR) is defined as how much the total number of papers a highly-cited researcher has published accounted for the total number of papers in the field. \( TP \) is the total number of papers in the field.

**Interdisciplinary Collaboration**

Boundary-crossing publication outside one’s own discipline represents a much more direct form of interdisciplinary information transfer than does citation, since information is presented by members of the discipline in which it originates (Pierce, 1999). So, in this study, 22 ESI fields of Thomson Reuters were used as a standard to classify publication, and we explored interdisciplinary collaboration behaviors of highly-cited scientists by analyzing their publication outside agricultural science.

**Cross-sector Collaboration**

Co-authorship of journal publications is often seen to represent the transfer of knowledge (Murray 2002). Two organizations are described as co-authors if they are both listed in the address filed of a paper. Authors listing more than one organizational affiliation on a paper thus represent collaboration between these organizations. These authors are seen as “boundary spanners” who facilitate the translation of knowledge between their organizational affiliations (Swan et al. 2007).

In order to determine the synergistic effect of university-industry-government collaboration, and find out a networked infrastructure for knowledge-based innovation systems, Leydesdorff has deduced an algorithm to measure the dynamic relations of “university-industry-government”, called TH algorithm (Leydesdorff, 2003). It is based on the mutual information in the three dimensions of the Triple Helix. In
information theory, Shannon defined entropy as probability of occurrence of discrete random events, which means the greater the uncertainty of events are, the greater the entropy is (Shannon, 1948). On the contrary, if a system operates safely, stably and orderly, then the value of entropy will be small. In the case of only one variable, the formula is:

\[ H = - \sum_j P_j \log_2 P_j \]

\( H \) represents entropy, that is, the average of information content, and \( P_j \) is the probability of the emergence of the number \( j \) information. Correspondingly, the entropy of two-dimensional distributed data is:

\[ H = - \sum_j \sum_l P_{ij} \log_2 P_{ij} \]

\( P_{ij} \) represents the joint probability distributions of event \( i \) and event \( j \). The mutual information between two dimensions of the probability distribution can be expressed as

\[ T_{ij} = H_i + H_j - H_{ij} \]. Suppose that \( u, j, g \) represent the number of publications of university, industry, and government, then the mutual information in three dimensions can be derived as

\[ T_{uig} = H_u + H_i + H_g - H_{ui} - H_{ug} - H_{ig} - H_{uig} \].

This study defines mutual information \( T \) as uncertainty of measurable variables, taking mutual information as a dynamic indicator to measure the close degree of each body. McGill pointed out that the three dimensional mutual information is the self-organizing measurement of the system, and is negatively correlated. It is used to explain the self-organizing problems of the network due to the lack of central coordination (McGill, 1954). When the value of \( T_{uig} \) is negative, it indicates that the three dimensional relationship plays the self-organization function in the system. As the value of \( T_{uig} \) is getting smaller, the whole system’s self-organization level will get higher, which suggests that the relation among UIG collaboration is getting closer.

According to Burt’s theory of structural holes (Ronald, 1992), nodes with high betweenness centrality play the role of a broker or gatekeeper to connect the nodes and sub-groups. So, they can most frequently control information flows in the network. Hence, in order to find which institutions occupied the dominant position in resource acquisition, we used the social network analysis method. We drew a UIG collaboration network map of highly-cited agricultural scientists’ affiliations, and ordered the networks nodes according to the value of betweenness centrality.

Data

Data Retrieval

In this study we selected the academic papers authored by 128 highly-cited agricultural scientists (2003-2013) according to 2015 World Most Influential Scientific Elites Report published by Thomson Reuters. Based on the report, publications authored by the 128 scientists and published from 2003 to 2013 are downloaded from SCI/SSCI of Thomson Reuters. Only document types of Articles
and Review are included. To guarantee the recall ratio, we retrieved data by authors’ family names and their first initials. After eliminating duplicates, 33760 publications were covered.

Data Cleaning

Author name disambiguation has become an active area of bibliometric research. It is a well-known problem in both cases: a single name may represent multiple individual authors and an individual author may publish or be indexed under multiple names, especially for none-English-speaking authors. In order to solve this problem, Thomson Reuters, Elsevier and other large database service providers introduce Research ID or Author ID as unique identifier of the author's academic identity. Some scholars have also developed a lot of methods for indentifying author’s name (Han et al., 2004; Han et al., 2005; Treeratpituk and Giles, 2009).

Thomson Reuters provided a lot of useful information about highly-cited scientists including their full names, affiliation names and Research IDs (but only 64 scientists’ Research ID can be found). For those scientists who lack Research ID information, this paper adopts K-way clustering algorithm proposed by Han (Han et al., 2004; Han et al., 2005) to identify highly-cited scientists’ papers. Finally, 12050 academic papers formed our target data set.

Data Normalization

Regionalization

In WoS dataset, C1 represents the author’s address, so we segment the address field so as to obtain the nation information. If this paper includes more than 2 national information, then it is considered as an international collaboration paper which reached 5324 pieces finally.

Subject Classification

This study is based on the Thomson Reuters ESI (Essential Science Indicators) classification system as standard. ESI subject classification mainly includes 22 major categories, which is classified according to the mapping relationship between the subject classification and the journal name. 11547 out of 12050 academic papers correspond to the correct ESI categories, and the other 503 need artificial judgment to match the ESI discipline classification system.

Author’s Affiliation Classification

Leydesdorff’s research result indicates that the mechanism type can be divided according to address cognitive word. This research’s data collecting process is as follows:

(a) We extracted institution addresses from 12050 papers data. Removing duplicate data, we got 27951 address information;

(b) Then, we began to identify and classify these addresses. If the address includes “UNIV OR COLL”, then it belongs to group “U” (university). Addresses that
Results and Analysis

General situation

Through the analysis of highly-cited scientists’ basic information provided by Thomson Reuters, we can find that 128 highly-cited agricultural scientists are from 26 countries/regions, and 85 institutions. (see Table 1). In the term of institution type, 95 highly-cited agricultural scientists come from colleges and universities, occupying 74.2% of the total; only one of highly-cited agricultural scientists works in enterprise (Dover Science Corporation in USA); the other 32 highly-cited agricultural scientists belong to government-led public institutions, such as United States Department of Agriculture (12), Spanish National Research Council (4), French Academy of Agriculture science (2), and so on.

<table>
<thead>
<tr>
<th>Countries</th>
<th>Number of highly-cited scientists</th>
<th>Institutions</th>
<th>Numbers of highly-cited scientists</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>40</td>
<td>USDA</td>
<td>12</td>
</tr>
<tr>
<td>UK</td>
<td>18</td>
<td>Univ Southampton</td>
<td>6</td>
</tr>
<tr>
<td>Germany</td>
<td>11</td>
<td>Cornell Univ.</td>
<td>5</td>
</tr>
<tr>
<td>Canada</td>
<td>7</td>
<td>Harvard Univ.</td>
<td>4</td>
</tr>
<tr>
<td>Australia</td>
<td>7</td>
<td>CSIC</td>
<td>4</td>
</tr>
<tr>
<td>Spain</td>
<td>5</td>
<td>Univ British Columbia</td>
<td>3</td>
</tr>
<tr>
<td>Italy</td>
<td>5</td>
<td>Wageningen Univ</td>
<td>3</td>
</tr>
<tr>
<td>France</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belgium</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkey</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>China (Taiwan, Hong Kong)</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Highly-cited agricultural scientists are productive author who publish an average of 94.1 papers in SCI/SSCI annually during the year of 2003-2013. Highly-cited agricultural scientists from developed countries, like the USA and European countries, lead a prominent performance in the output of their academic publications (Table 2).
Table 2 Publication of Top 10 Highly-cited Agricultural Scientists

<table>
<thead>
<tr>
<th>Names</th>
<th>Nations</th>
<th>Affiliations</th>
<th>Citations</th>
<th>Papers</th>
<th>Average Citations</th>
<th>Highly-cited Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frank B Hu</td>
<td>USA</td>
<td>Harvard Univ.</td>
<td>51071</td>
<td>551</td>
<td>92.7</td>
<td>70</td>
</tr>
<tr>
<td>Walter C Willett</td>
<td>USA</td>
<td>Harvard Univ.</td>
<td>48721</td>
<td>635</td>
<td>79.7</td>
<td>62</td>
</tr>
<tr>
<td>David Julian McClements</td>
<td>USA</td>
<td>University of Massachusetts</td>
<td>14908</td>
<td>342</td>
<td>43.6</td>
<td>33</td>
</tr>
<tr>
<td>Rattan Lal</td>
<td>USA</td>
<td>University at Columbus</td>
<td>12621</td>
<td>274</td>
<td>46.1</td>
<td>11</td>
</tr>
<tr>
<td>Bruno Vellas</td>
<td>France</td>
<td>Université de Toulouse</td>
<td>12595</td>
<td>315</td>
<td>40.0</td>
<td>11</td>
</tr>
<tr>
<td>Barry M Popkin</td>
<td>USA</td>
<td>University of North Carolina</td>
<td>12192</td>
<td>172</td>
<td>70.9</td>
<td>18</td>
</tr>
<tr>
<td>Philip C Galder</td>
<td>UK</td>
<td>University of Southampton</td>
<td>11503</td>
<td>203</td>
<td>56.7</td>
<td>18</td>
</tr>
<tr>
<td>Mike E Goddard</td>
<td>Australia</td>
<td>The University of Melbourne</td>
<td>11336</td>
<td>124</td>
<td>91.4</td>
<td>18</td>
</tr>
<tr>
<td>Guoyao Wu</td>
<td>USA</td>
<td>Texas A&amp;M University</td>
<td>11329</td>
<td>204</td>
<td>55.5</td>
<td>13</td>
</tr>
<tr>
<td>R Paul Ross</td>
<td>Ireland</td>
<td>Teagasc</td>
<td>10509</td>
<td>274</td>
<td>38.4</td>
<td>12</td>
</tr>
</tbody>
</table>

International collaboration

From 2003 to 2013, highly-cited agricultural scientists have published 5324 international collaboration papers in the field of agriculture science, accounting for 44.2% of the total number of papers. As Fig. 2 shows, the output and the proportion of international collaboration papers present a year-by-year growth trend. Apparently, in 2012 and 2013, the proportion of international collaboration papers from highly-cited agricultural scientists is more than 50% of the total. It has become an obvious trend that highly-cited agricultural scientist are paying more attention to international collaboration.

Fig. 2 International Collaboration Papers of Highly-cited Agricultural Scientists from 2003 to 2013

ICAI indicator can reflect the collaboration degree of highly-cited scientists in multi-dimension. Table 3 shows some relative information about active agricultural...
scientists. The results show that highly-cited agricultural scientists in developed countries, such as the United States and European countries, are leading a more active participation in international collaboration. They continue to improve their own research strength by expanding international collaboration chances and using global scientific resources effectively. Meanwhile, two agricultural scientists from China and Thailand have a higher participation degree in the international collaboration, indicating that developing countries are also seeking international collaboration opportunities and strengthening their ability, in order to make up insufficiencies of science and technology resources in their countries.

Table 3 ICAI of Top 10 Highly-cited Agricultural Scientists

<table>
<thead>
<tr>
<th>Names</th>
<th>Nations</th>
<th>Affiliations</th>
<th>ICAI</th>
<th>ANCP</th>
<th>PCR</th>
<th>CLI</th>
<th>ICR</th>
<th>PPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rajeev K Varshney</td>
<td>Australia</td>
<td>Univ Western Australia</td>
<td>0.192</td>
<td>1.148</td>
<td>0.822</td>
<td>0.550</td>
<td>33</td>
<td>0.011</td>
</tr>
<tr>
<td>Soottawat Benjakul</td>
<td>Thailand</td>
<td>Prince Songkla Univ</td>
<td>0.084</td>
<td>0.384</td>
<td>0.549</td>
<td>0.604</td>
<td>22</td>
<td>0.030</td>
</tr>
<tr>
<td>Pete Smith</td>
<td>UK</td>
<td>Univ Aberdeen</td>
<td>0.051</td>
<td>0.574</td>
<td>0.822</td>
<td>0.151</td>
<td>51</td>
<td>0.014</td>
</tr>
<tr>
<td>Gary Williamson</td>
<td>UK</td>
<td>Univ Leeds</td>
<td>0.033</td>
<td>0.379</td>
<td>0.729</td>
<td>0.392</td>
<td>26</td>
<td>0.012</td>
</tr>
<tr>
<td>David L Jones</td>
<td>UK</td>
<td>Bangor Univ</td>
<td>0.026</td>
<td>0.249</td>
<td>0.732</td>
<td>0.244</td>
<td>33</td>
<td>0.018</td>
</tr>
<tr>
<td>Maurizio Battino</td>
<td>Italy</td>
<td>Univ Politecn Marche</td>
<td>0.025</td>
<td>0.506</td>
<td>0.662</td>
<td>0.471</td>
<td>25</td>
<td>0.006</td>
</tr>
<tr>
<td>Johannes Lehmann</td>
<td>USA</td>
<td>Cornell Univ</td>
<td>0.023</td>
<td>0.318</td>
<td>0.783</td>
<td>0.257</td>
<td>33</td>
<td>0.011</td>
</tr>
<tr>
<td>Alan Crozier</td>
<td>UK</td>
<td>Univ Glasgow</td>
<td>0.017</td>
<td>0.411</td>
<td>0.695</td>
<td>0.470</td>
<td>16</td>
<td>0.008</td>
</tr>
<tr>
<td>Steven J Lehotay</td>
<td>USA</td>
<td>USDA</td>
<td>0.017</td>
<td>0.582</td>
<td>0.636</td>
<td>0.657</td>
<td>15</td>
<td>0.005</td>
</tr>
<tr>
<td>Yulong Yin</td>
<td>China</td>
<td>Chinese Acad Sci</td>
<td>0.015</td>
<td>0.399</td>
<td>0.626</td>
<td>0.549</td>
<td>8</td>
<td>0.014</td>
</tr>
</tbody>
</table>

In addition, based on the five international collaboration indicators (ANCP, PCR, CLI, ICR, PPR), we classified 128 highly-cited agricultural scientists by using hierarchical cluster analysis. According to Table 4, 31 scientists in cluster 1, occupying 24.2%, perform more actively in the international collaboration. 61 scientists in cluster 2, accounting for 47.7%, have a higher ICR, but lower CLI and PPR than scientists in cluster 1. 36 scientists in cluster 3, occupying 28.1%, get lower grades in the 5 indicators than scientists in the other two clusters, so it can be regarded as relatively inactive in the international collaboration.

Table 4 Cluster of International collaboration activity index

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Numbers of scientists</th>
<th>Average of ICAI</th>
<th>Average of ANCP</th>
<th>Average of PCR</th>
<th>Average of CLI</th>
<th>Average of ICR</th>
<th>Average of PPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31</td>
<td>0.016</td>
<td>0.262</td>
<td>0.427</td>
<td>0.496</td>
<td>14.5</td>
<td>0.013</td>
</tr>
<tr>
<td>2</td>
<td>61</td>
<td>0.004</td>
<td>0.121</td>
<td>0.549</td>
<td>0.156</td>
<td>21.1</td>
<td>0.009</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
<td>0.001</td>
<td>0.037</td>
<td>0.209</td>
<td>0.141</td>
<td>9.4</td>
<td>0.005</td>
</tr>
</tbody>
</table>

The method of hierarchical clustering: using Ward algorithm to cluster, and choosing Squared Euclidean distance as the metric standards, as well as, Z score to eliminate the influence of the dimension of each index.
Interdisciplinary Collaboration

From 2003 to 2013, highly-cited agricultural scientists have published 9006 co-authored papers, accounting for 74.7%. Among them, the number of interdisciplinary collaboration papers is 3691, which take up 30.6% of the total. Besides, there are up to 5315 interdisciplinary collaboration papers, occupying 44.1%, as shown in Fig. 1. We can see that only 3 out of 128 scientists published papers in the single subject, indicating interdisciplinary collaboration is a common phenomenon in highly-cited agricultural scientists group.

As Table 5 shows, interdisciplinary collaboration papers in the field of agriculture science are mainly concentrated on neighbor areas, such as biology and biochemistry, animal and plant sciences, and chemistry. On the contrary, few highly-cited agricultural scientists publish papers in long-distance discipline areas, such as material science, economics and business, space science.

Table 5 Publication Fields of Highly-cited Agricultural Scientists

<table>
<thead>
<tr>
<th>Fields</th>
<th>Numbers of scientists</th>
<th>Proportion of scientists (%)</th>
<th>Output (papers)</th>
<th>Publication proportion (%)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology and Biochemistry</td>
<td>80</td>
<td>62.50</td>
<td>744</td>
<td>8.26</td>
</tr>
<tr>
<td>Plant and animal science</td>
<td>66</td>
<td>51.56</td>
<td>565</td>
<td>6.27</td>
</tr>
<tr>
<td>Chemistry</td>
<td>55</td>
<td>42.97</td>
<td>313</td>
<td>3.47</td>
</tr>
<tr>
<td>Clinical medicine</td>
<td>55</td>
<td>42.97</td>
<td>1358</td>
<td>15.07</td>
</tr>
<tr>
<td>Environmental Science and ecology</td>
<td>50</td>
<td>39.06</td>
<td>650</td>
<td>7.21</td>
</tr>
<tr>
<td>Pharmacology and Virology</td>
<td>46</td>
<td>35.94</td>
<td>221</td>
<td>2.45</td>
</tr>
<tr>
<td>Molecular biology and genetics</td>
<td>42</td>
<td>32.81</td>
<td>252</td>
<td>2.80</td>
</tr>
<tr>
<td>Microbiology</td>
<td>28</td>
<td>21.88</td>
<td>163</td>
<td>1.81</td>
</tr>
<tr>
<td>Earth Sciences</td>
<td>26</td>
<td>20.31</td>
<td>249</td>
<td>2.76</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>26</td>
<td>20.31</td>
<td>278</td>
<td>3.09</td>
</tr>
<tr>
<td>Neuroscience and praxiology</td>
<td>19</td>
<td>14.84</td>
<td>157</td>
<td>1.74</td>
</tr>
<tr>
<td>Engineering</td>
<td>17</td>
<td>13.28</td>
<td>54</td>
<td>0.60</td>
</tr>
<tr>
<td>Computer science</td>
<td>15</td>
<td>11.72</td>
<td>23</td>
<td>0.26</td>
</tr>
<tr>
<td>Psychiatry and Psychology</td>
<td>12</td>
<td>9.38</td>
<td>35</td>
<td>0.39</td>
</tr>
</tbody>
</table>
The UIG collaboration degree of highly-cited agricultural scientists presented a volatile growth trend from 2003 to 2013 (Fig. 4). According to affiliation types, universities (-54.17) owned higher UIG (inverse indictor) than governments (-84.17) and industries (-198.24) on average, which illustrates that the UIG collaboration degree of highly-cited agricultural scientists from universities is lower than the scientists from the other two types of institutions. The reason may be that colleges and universities are usually more inclined to basic research work, but the government and enterprises spend a lot on the commercialization of research findings. However, with universities paying more attention to the commercialization of research achievements, we can foresee that the gap of the UIG collaboration degree between universities and the other two types of institutions will be narrowed.

![Fig. 4 T_UIG of Highly-cited Agricultural Scientists from 2003 to 2013](image)

According to the top 10 highly-cited agricultural scientists in UIG collaboration, the majority of collaboration institutions came from their own countries. Some of the world famous transnational enterprises, such as Mars Company and Pfizer, have become some highly-cited researchers’ important UIG collaboration partners.

Table 6 Top 10 Highly-cited researchers’ information of UIG

<table>
<thead>
<tr>
<th>Names</th>
<th>Affiliations (categories)</th>
<th>( T_{UIG} )</th>
<th>UIG Collaboration institutions* (categories)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boxin Ou</td>
<td>Dover Sciences (I)</td>
<td>-198.24</td>
<td>Natl Univ Singapore (U); USDA (G)</td>
</tr>
<tr>
<td>Gary R Beecher</td>
<td>USDA (G)</td>
<td>-194.32</td>
<td>Tufts Univ (U); Mars Inc (I)</td>
</tr>
<tr>
<td>Bruno Vellas</td>
<td>CHU Toulouse (U)</td>
<td>-179.37</td>
<td>Pfizer Ltd (I); INSERM (G)</td>
</tr>
<tr>
<td>Liwei Gu</td>
<td>Univ Florida (U)</td>
<td>-172.04</td>
<td>Ocean Spray Cranberries Inc (I); USDA (G)</td>
</tr>
<tr>
<td>Joanne M Holden</td>
<td>Beltsville Human Nutr Res Ctr (G)</td>
<td>-162.83</td>
<td>Virginia Polytech Inst &amp; State Univ (U); Mars Inc (I)</td>
</tr>
</tbody>
</table>
As the scientific collaboration network of highly-cited researchers displays (Fig. 5), most of the universities are located in the network’s core position with 0.671 betweenness centrality on average. So we can figure out that collaboration between universities is still an important form for scientific collaboration. Nevertheless, some government-dominated public institutions appear at the edge of scientific collaboration networks, and can function as bridge (the betweeness centrality of these 6 governmental public institutions is 0.655 on average). And they build up UIG collaboration relationships with some universities. Besides, the Swiss Nestle (betweeness centrality value: 0.539) also accounts for a place in the network map. It co-authored with 34 scientists, and published a total of 135 academic papers, illustrating this corporation’s imperative status in the relation of UIG collaboration of this field.

**Fig. 5 Scientific Collaboration Network of Highly-cited Agricultural Scientists**

Network setting parameters and algorithms: the size of the nodes denotes the number of the publications of the institution (≥20); the color depth of the nodes denotes the value of betweenness centrality; the thickness of edges denotes the number of collaboration papers. Force Altlas algorithm is used for layout. Label color represents UIG: Black=U, Red=I, Blue=G).
Conclusion and Discussion

This study tried to contribute some information to the understudied field by making a bibliometric analysis on academic papers authored by 128 highly-cited agricultural scientists, based on the analysis framework of cross-boundary collaboration from three dimensions: international collaboration, interdisciplinary collaboration and university-industry-government (UIG) collaboration. The results showed that, firstly, Colleges and universities were the main affiliations of the highly-cited agricultural scientists, and highly-cited scientists from developed countries performed outstanding in academic paper output. Secondly, international collaboration papers of highly-cited agricultural scientists showed an increasing trend. 71.9% of the highly-cited agricultural scientists are active in the international collaboration. Thirdly, interdisciplinary collaboration among highly-cited agricultural scientists was very common. In particularly, adjacent subjects became the main field of interdisciplinary collaboration of highly-cited agricultural scientists. On the contrary, few highly-cited agricultural scientists published papers in long-distance discipline areas. Fourthly, UIG collaboration of highly-cited agricultural scientists from universities was less active than those from government institutions and enterprises. Nestle has become an important partner of highly-cited agricultural scientists in UIG collaboration.

For training a new generation of cross-boundary scientists, it is very important to design new educational strategies. Within university settings, courses and research mentorship opportunities emphasizing transdisciplinary approaches to science and community action should be developed (Stokols, 2006). And within community contexts, new workshops and training modules focusing on the challenges associated with cross-boundary collaborations and practical strategies that can be used by participants to improve their communications and coordination efforts would provide a valuable resource for the members of scientific teams, community coalitions, and inter-sector partnerships. Meanwhile, the knowledge innovation of universities should break through itself, paying attention to collaboration with the government and enterprises to promote heterogeneous resources flow in universities, government and enterprises (Becher and Trowler, 2001).

Bibliometric studies will help research organizations acquire a more comprehensive understanding of cross-boundary collaboration characteristics of high-cited scientists. The analysis of the characteristics of research collaboration is also very relevant to policymaking. For example, funding agencies and decision makers can allocate budgetary support for collaboration between university and industry and the facilitation of multidisciplinary communication.

Acknowledgments

This work was financially supported by the ISTIC-ELSEVIER Journal Assessment Research Center Fund and NSFC (71373252).

Reference:

B. L. Clarke, Communication Patterns of Biomedical Scientists, Federation Proceedings 26 (1967) 1288-1292.


853


A scientometric method for assessing an institution’s scientific collaboration policy

Hamid Bouabid,
Faculty of Science, Mohammed V University in Rabat, 4, Avenue Ibn Battuta, BP 1014 RP, Rabat, Morocco, h.bouabid@hotmail.com

Abstract
The paper suggests a comprehensive method for an institution's scientific collaboration policy assessment. This method is based mainly on indicators: signed agreements, co-publications, their citations and scientific fields as front science. The method also uses a mapping of collaborative institutions and fields to assess the proximity of the institution in its collaborative cluster as well as multidisciplinarity in this cluster. Three major classes of collaborative institutions are derived from the method: (1) a class where there is coherence between co-publications and agreements; the university has to faithfully intensify its collaboration with this class; (2) a class where there is a substantial amount of co-publications but no collaboration agreement; the university should consider setting up a formal frame for collaborative research; and (3) a class grouping institutions which cited the university knowledge and considered as potential partners for the university.

Conference topic
Methods and techniques - Science policy and research assessment.

Introduction and background
Science and technology (S&T) are increasingly open, multidisciplinary and highly integrative. Therefore, collaboration offers several opportunities and contributes to increased productivity and excellence in S&T. Several authors have demonstrated that there is a strong correlation between scientific collaboration and research productivity and impact (Katz and Martin, 1997, Lee and Bozeman, 2005, He et al., 2009, Abramo et al., 2011a and Finlay et al., 2012). Defazio et al. (2009) modeled the effect of collaboration and funding on productivity for three periods: pre-funding, during funding and post-funding. They found that receiving funding increases researcher productivity by approximately 14%, while collaborating with a partner in the network in the post-funding period increases productivity by approximately 70%. The positive
effect of collaborations on the impact of papers has also been proven using citations or impact (Katz and Hicks, 1997, Guerrero-Bote et al., 2013, He et al., 2009, Levitt and Thelwell, 2010). While there exists a rich literature offering findings and conclusions on collaboration assessments at the scientist -micro- level (Traore and Landry, 1997; Melin, 2000; Hara et al., 2003; Jha and Welch, 2010; Lee and Bozeman, 2005), or at the regional, national or international -macro- level (Tijis et al., 2010; Hoekman et al, 2010; Choi, 2012, Bordons et al., 2013), there are far fewer assessments of scientific collaboration at the institutional level, as done by Adams (2005), Ponds et al. (2007), Ortega and Aguillo (2013), Barth et al. (2014), Han et al. (2014) and Yan and Guns (2015). Furthermore, the methods applied in these papers are different and less methodologically organized to link collaborative inputs (agreements) to outputs (co-publications) for policy issues. Indeed, for example, Ponds et al. (2007) make use of the gravity model to examine the proximity effect of collaboration between different kinds or similar institutions. The work by Barth et al. (2014) was based on network and cluster analysis (carried out using VOSviewer software) to identify similar institutions for collaboration and potential partner’ ones. Finally, Yan and Guns (2015) applied eight algorithms to collaboration at the author, institution and country levels for predicting and shaping the dynamics of collaboration starting from existing topology.

**Data and Method**

In this assessment, all frameworks, agreements, conventions and Memorandum of Understanding (MoU), were considered acts of collaboration. Other types of contracts, tenders, and sponsoring agreements or similar were excluded. During the whole period, UM5S signed 248 agreements.

The University Mohammed V - Souissi (UM5S) comprises the following 10 institutions:

- Medicine and pharmacy faculty;
- Dental medicine faculty;
- National high school of technology teaching;
- National high school of computer science and systems analysis;
- Law and Economics faculty in Souissi;
- Law and Economics faculty in Sale;
- Education sciences faculty;
- Institute for African studies;
- Institute for Arabic studies and research;
- University Institute for scientific research.

The Web of Science database was used to retrieve publications and co-publications (co-authored ones), where all kinds of publications were considered (article, review, proceeding, letter, editorial, etc). In order to compare agreements and co-publications, the analysis covers
the second period from 2010 to 2013 (end of June 2013). UM5S published almost 600 publications during this period, of which 63% were co-publications. In this analysis, a co-publication is defined as a paper authored by the UM5S and at least one other institution (national or international).

To map and analyze co-publications, Salton's Cosine (Salton, 1983) is used as the index for calculating the proximity between the entities (institution, field, key-words, etc) of the network. Figure 1 reports the proposed method to assess the policy of university scientific collaboration.

![Fig. 1: Scheme and stream of the proposed method for collaboration policy assessment](image)

Figure 1 plots the method process. The first stage consists of compiling data and information as a policy inputs. The second stage refers to the process of the analysis itself comprising three levels: (i) the stock of active collaboration agreements, which represents, for a given year, the difference between the number of cumulative valid agreements and the number of expired agreements up to that year, (ii) the coherence between agreements and co-publications and (iii) the impact of co-publications as measured by their citations. To achieve this analysis five indicators are produced in order to build up the three classes of collaborative institutions. In another institution's assessment, this number will depend on its collaboration policy.

**Results and discussion**

For the whole period, the stock of active collaboration agreements has been calculated, as described in the Data and Method section. Its analysis is not part of this paper.
**Coherence between agreements and co-publications**

Through the indicator of co-publication, the method quantifies the share of each institution in the collaboration activities of the university. For UM5S, the method has shown a national co-publication dominance amongst overall university co-publications (Table 1). Major national collaborative institutions are the Specialized Hospitals in Rabat, Hassan II University in Casablanca, and Ibn Tofail University in Kenitra. The different institutions belonging to the CHU Ibn Sina are from now on not considered separately but included in it. Furthermore, Academy Hassan II of Science and Technology is excluded since it is more of a funding body and not a research institution. Similarly, and for coherence purposes, the Ministry of Health will not be considered at all as a collaborative institution.

On the international level it is also noted that there is a diversity of collaborative institutions (Table 1). It is worth mentioning that two-thirds of the 39 UM5S collaborative institutions (above the threshold of 10 co-publications) according to co-publications with UM5S are not under any collaboration agreement with UM5S, demonstrating a clear distortion between agreement as input proxy and co-publication as its output proxy. Cross-checking the results in Table 1 from either signed agreements and co-publications according to the method in Figure 1 allows us to build up two classes:

- **Class 1**: institutions with which the university has collaboration agreements as well as co-publications with their researchers, such as National Institute of Health, University Mohammed V Agdal, Ibn Sina University Hospital (CHU), University Ibn Tofail, CNRS-France, University Zaragoza, University Bordeaux, Université de Montréal, Université Laval, Cheikh Zaid Hospital, University Cadi Ayyad;

- **Class 2**: institutions with which there is no collaboration agreement, but there is substantial co-publication activity, such as: Military University Hospital Mohammed V, University Hassan II (Casablanca and Mohammedia), University Autonoma of Madrid, King Saud University, University Mohammed I, University of Technology Czestochowa, University Lodz, University of Kyoto, European Hospital Georges Pompidou, University Nancy 1, Hospital Virgen Camino, Medical Center Erasmus, University Valencia, University Rouen, CHU Limoges, University Liege, CHU of Tours, University Sidi Mohammed Ben Abdellah, University Paris 7.

While Autonoma University in Spain appeared to be a collaborative institution to the UM5 without any agreement, we found that the university had signed several agreements with other Spanish institutions but with no scientific output, such as: University Laguna, Universia
Holding, university Murcia, university Las Islas Baléares and University Cadiz. It is one of the illustrative cases of discrepancy between agreements and co-publications.

### Table 1: Breakdown of the UM5S's collaborative institutions (with a threshold of 10 co-publications)

<table>
<thead>
<tr>
<th>Institution</th>
<th>Agreement</th>
<th>Number of Co-publications</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natl Inst Health</td>
<td>x</td>
<td>31</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Mohammed v agdal</td>
<td>x</td>
<td>31</td>
<td>Morocco</td>
</tr>
<tr>
<td>Ibn Sina University Hosp*</td>
<td>x</td>
<td>29</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Ibn Tofail</td>
<td>x</td>
<td>26</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Hassan II</td>
<td></td>
<td>26</td>
<td>Morocco</td>
</tr>
<tr>
<td>Military Hosp Univ Med V</td>
<td></td>
<td>25</td>
<td>Morocco</td>
</tr>
<tr>
<td>Hassan II Acad Sci &amp; Technol</td>
<td></td>
<td>22</td>
<td>Morocco</td>
</tr>
<tr>
<td>Hop Enfants Rabat*</td>
<td></td>
<td>19</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Autonoma Madrid</td>
<td></td>
<td>14</td>
<td>Spain</td>
</tr>
<tr>
<td>Hop Specialites *</td>
<td></td>
<td>14</td>
<td>Morocco</td>
</tr>
<tr>
<td>CNRS</td>
<td>x</td>
<td>14</td>
<td>France</td>
</tr>
<tr>
<td>King Saud University</td>
<td></td>
<td>14</td>
<td>Saudi Arabia</td>
</tr>
<tr>
<td>University Mohamed I</td>
<td></td>
<td>14</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Zaragoza</td>
<td>x</td>
<td>14</td>
<td>Spain</td>
</tr>
<tr>
<td>El Ayachi University Hosp*</td>
<td></td>
<td>13</td>
<td>Morocco</td>
</tr>
<tr>
<td>Czestochowa Technology University</td>
<td></td>
<td>12</td>
<td>Poland</td>
</tr>
<tr>
<td>University Lodz</td>
<td></td>
<td>12</td>
<td>Poland</td>
</tr>
<tr>
<td>Kyoto university</td>
<td></td>
<td>11</td>
<td>Japan</td>
</tr>
<tr>
<td>University Bordeaux</td>
<td>x**</td>
<td>11</td>
<td>France</td>
</tr>
<tr>
<td>University Montreal</td>
<td>x</td>
<td>11</td>
<td>Canada</td>
</tr>
<tr>
<td>University Laval</td>
<td>x</td>
<td>11</td>
<td>Canada</td>
</tr>
<tr>
<td>Minist Health</td>
<td>x</td>
<td>11</td>
<td>Morocco</td>
</tr>
<tr>
<td>Hop Cheikh Zaid</td>
<td>x</td>
<td>11</td>
<td>Morocco</td>
</tr>
<tr>
<td>Hop Europeen Georges Pompidou</td>
<td></td>
<td>10</td>
<td>France</td>
</tr>
<tr>
<td>University Nancy 1</td>
<td></td>
<td>10</td>
<td>France</td>
</tr>
<tr>
<td>Hosp Virgen Camino</td>
<td></td>
<td>10</td>
<td>Spain</td>
</tr>
<tr>
<td>Erasmus Med Center</td>
<td></td>
<td>10</td>
<td>Netherlands</td>
</tr>
<tr>
<td>University Valencia</td>
<td></td>
<td>10</td>
<td>Spain</td>
</tr>
<tr>
<td>University Rouen</td>
<td></td>
<td>10</td>
<td>France</td>
</tr>
<tr>
<td>CHU Limoges</td>
<td></td>
<td>10</td>
<td>France</td>
</tr>
<tr>
<td>University Liege</td>
<td></td>
<td>10</td>
<td>Belgium</td>
</tr>
<tr>
<td>Hop Specialities Rabat*</td>
<td></td>
<td>10</td>
<td>Morocco</td>
</tr>
<tr>
<td>CHU Tours</td>
<td></td>
<td>10</td>
<td>France</td>
</tr>
<tr>
<td>Inst Natl Oncol*</td>
<td></td>
<td>10</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Cadi Ayyad</td>
<td>x</td>
<td>10</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Sidi Mohamed Ben Abdellah</td>
<td></td>
<td>10</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Paris 07</td>
<td></td>
<td>10</td>
<td>France</td>
</tr>
</tbody>
</table>

* Ibn Sina Hosp institutions

** including ENSEIRB, Univ Michel de Montaigne and univ Montesquieu
The same case holds for some French institutions. Indeed, while no agreement was signed with University Paris 7 despite many co-publications, we found that several agreements were set with institutions in Paris such as: University Paris 13, University Paris Dauphine, University Paris René Descartes, Institute for Languages and Civilizations.

**Impact of co-publications**

This part of the method is mainly based on science mapping which has been proven to be a useful tool in assessment.

1. **Impact of co-publications**

Collaboration with high-productivity partners tends to increase impact as proved by Katz and Hicks (1997) who found that while collaborating with an author from the home or domestic institution increases the average impact by approximately 0.75 citations, collaborating with an author from a foreign institution increases the impact by about 1.6 citations. Likewise, Abramo et al. (2011b) confirmed that more productive scientists and those with high impact collaborate more abroad than their colleagues (They warned, however, that the reverse was not always true). In this regard, we provide in Table 2 the impact indicator for institutions of the two classes in order to ascertain those with higher impact than that of UM5S. Table 2 shows that almost all of the collaborative institutions were more knowledge productive (not normalized to university size). Despite its slightly lower impact rate, but closer to that of UM5S, Ibn Tofail University is one of the small-size but performed universities in Morocco (Bouabid, 2014).

Three institutions were excluded for having a number of publications below the threshold of 100 publications during the same period: National Institute of Health (Class 1), Hospital Cheikh Zaid (Class 1) and Hospital Virgen Camino (Class 2). Two other institutions were excluded for having lower impact than that of UM5S (shown in grey in Table 2).

At the end of this stage of the method's stream, the two classes comprise the following:

- **Class 1**: University University Mohammed V-Agdal, University Ibn Tofail, CNRS-France, University Zaragoza, University Bordeaux, Université de Montréal, Université Laval, University Cadi Ayyad. The scientific collaboration with this class should be strengthened since there is correlation between co-publication practices and collaboration agreements;

- **Class 2**: University Hassan II (Casablanca and Mohammedia), University Autonoma of Madrid, King Saud University, University Mohammed I, University of Technology Czestochowa, University Lodz, University of Kyoto, European Hospital Georges Pompidou, University Nancy 1, Medical Center Erasmus, University Valencia, University
Rouen, University Liege, CHU of Tours, University Sidi Mohammed Ben Abdellah, University Paris 7. The scientific collaboration for this class is distorted and unbalanced and the university should consider setting up a formal framework for collaboration for its researchers to enhance their research activities.

Table 2: UM5S’s collaborative institutions with numbers of publications during the same period of study (with a threshold of 100 publications), citations and impact (citations/publication)

<table>
<thead>
<tr>
<th>Institution</th>
<th>Nb. Publications (p)</th>
<th>Nb. Citations (c)</th>
<th>Impact (c/p)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Class 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Natl Inst Health</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Mohammed V-Agdal</td>
<td>799</td>
<td>4,477</td>
<td>5.60</td>
</tr>
<tr>
<td>Ibn Sina University Hosp</td>
<td>454</td>
<td>690</td>
<td>1.52</td>
</tr>
<tr>
<td>Ibn Tofail</td>
<td>328</td>
<td>592</td>
<td>1.80</td>
</tr>
<tr>
<td>CNRS</td>
<td>139,314</td>
<td>299,263</td>
<td>2.15</td>
</tr>
<tr>
<td>Zaragoza</td>
<td>7,396</td>
<td>38,321</td>
<td>5.18</td>
</tr>
<tr>
<td>Bordeaux</td>
<td>10,221</td>
<td>72,763</td>
<td>7.12</td>
</tr>
<tr>
<td>Montreal</td>
<td>21,626</td>
<td>139,397</td>
<td>6.45</td>
</tr>
<tr>
<td>Laval</td>
<td>11,461</td>
<td>70,534</td>
<td>6.15</td>
</tr>
<tr>
<td>Hosp Cheikh Zaid</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Cadi Ayyad</td>
<td>1,289</td>
<td>7,981</td>
<td>6.19</td>
</tr>
<tr>
<td><strong>Class 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Military Hosp Univ Med V</td>
<td>235</td>
<td>211</td>
<td>0.90</td>
</tr>
<tr>
<td>Hassan II</td>
<td>773</td>
<td>7,237</td>
<td>9.36</td>
</tr>
<tr>
<td>Antonoma Madrid</td>
<td>9,284</td>
<td>67,624</td>
<td>7.28</td>
</tr>
<tr>
<td>King Saud</td>
<td>10,399</td>
<td>43,073</td>
<td>4.14</td>
</tr>
<tr>
<td>Mohammed I</td>
<td>492</td>
<td>6,654</td>
<td>13.52</td>
</tr>
<tr>
<td>Czestochowa</td>
<td>1,232</td>
<td>2,469</td>
<td>2.00</td>
</tr>
<tr>
<td>Lodz</td>
<td>7,056</td>
<td>27,089</td>
<td>3.84</td>
</tr>
<tr>
<td>Kyoto</td>
<td>28,153</td>
<td>170,073</td>
<td>6.04</td>
</tr>
<tr>
<td>Hop European Georges Pompidou</td>
<td>288</td>
<td>2,786</td>
<td>9.67</td>
</tr>
<tr>
<td>Nancy 1</td>
<td>978</td>
<td>7,611</td>
<td>7.78</td>
</tr>
<tr>
<td>Hosp Virgen Camino</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Med Cntr Erasmus</td>
<td>20,636</td>
<td>162,426</td>
<td>7.87</td>
</tr>
<tr>
<td>Valencia</td>
<td>11,416</td>
<td>98,050</td>
<td>8.59</td>
</tr>
<tr>
<td>Rouen</td>
<td>2,704</td>
<td>18,644</td>
<td>6.89</td>
</tr>
<tr>
<td>CHU Limoges</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Liege</td>
<td>8,536</td>
<td>66,025</td>
<td>7.73</td>
</tr>
<tr>
<td>CHU Tours</td>
<td>482</td>
<td>2,940</td>
<td>6.10</td>
</tr>
<tr>
<td>Sidi Mohamed Ben Abdellah</td>
<td>309</td>
<td>2,635</td>
<td>8.53</td>
</tr>
<tr>
<td>Paris 7</td>
<td>16,099</td>
<td>163,552</td>
<td>10.16</td>
</tr>
<tr>
<td><strong>UM5S</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UM5S</td>
<td>596</td>
<td>1,085</td>
<td>1.82</td>
</tr>
<tr>
<td><strong>Morocco</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morocco</td>
<td>8,542</td>
<td>25,053</td>
<td>2.93</td>
</tr>
</tbody>
</table>

* p: is the number of publications during the period from January 1st 2010 to June 30th 2013.
** c: is the number of citations gained by the publications p in the time window: January 1st 2010 to December 31th 2014.
-- : means that the number of publications is below the threshold
As a policy implication, for the Class 2 institutions a 'bottom-up' approach is required to identify university researchers involved who would contribute to setting up formal scientific collaboration with these institutions. In fact, the scientific collaborations were found to be mostly based on individual initiative as was reported by Bordons et al. (2013) rather than a result of policy action. Melin (2000) suggested that scientists themselves should choose with whom they would like to cooperate, and under which forms.

2. Multidisciplinarity in co-publications

To assess proximity to collaborative institutions and the multidisciplinarity in collaboration of UM5S, as a key component of the method clusters are drawn in a heterogeneous map crossing both collaborating institutions and scientific fields for the university in terms of co-publications (Figure 2). Despite its relative complexity, this kind of map provides a complete visualization and is of greater value by combining both institutions and fields. For the first point, higher proximity - e.g. closeness - is shown to be more prevalent with specialized institutions in Rabat such as Children’s Hospital (Hôpital des enfants), Ibn Sina University Hospital, Mohammed V Military Hospital and National Institute of Health, as well as with Ibn Tofail University (35 kilometers north of Rabat). For the second point, Figure 2 pictures a multidisciplinary collaboration with very limited institutions: Ibn Tofail University (Class 1) and Georges Pompidou European Hospital (Class 2) as identified by the overlap of more than two different circles - sub-clusters (marked with arrows in Fig. 2). Policy collaboration focus should be on these institutions as multidisciplinary collaborative ones.

3. Research Fronts/advanced knowledge in co-publications

In addition to the classification presented above, a qualitative indicator is employed to check the relevance of considering these classes in the university collaboration policy. The objective is to ensure that the topics addressed in the research collaboration with these institutions are in 'Research Fronts' as defined by Thomson Reuters in its Science Watch©. This part of analysis is not presented in this paper. However, it was found that the topics addressed in the scientific collaboration were generally in the Research Fronts such as: genetics, genomes and gene expression, tissue engineering, immunology, aging, thin films, remote sensing, neurosciences, polymer, etc.
Finally, considering the overall knowledge production stream, the results showed that some institutions from emerging countries such as China, India, Brazil, and Turkey are likely to be potential collaborative partners for the university since they have been citing its papers (Table 3) and with respect to their fast-growing knowledge production. These institutions are grouped into a third class as potential partners. Surprisingly, no co-publication or agreement were found with any of these institutions. Only institutions having more than five citations each (as a threshold) to UM5S’ publications are considered as citing institution from these countries.
Table 3: Breakdown by country's collaborative institutions of the number of signed agreements, co-publications and received citations (for the 15 first countries and in bold the first five scores)

<table>
<thead>
<tr>
<th>Country's collaborative institutions</th>
<th>Co-publications</th>
<th>Agreements</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Rank</td>
<td>Number</td>
</tr>
<tr>
<td>Morocco</td>
<td>743</td>
<td>1</td>
<td>118</td>
</tr>
<tr>
<td>France</td>
<td>114</td>
<td>2</td>
<td>52</td>
</tr>
<tr>
<td>Spain</td>
<td>33</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>USA</td>
<td>19</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Japan</td>
<td>19</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>17</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Canada</td>
<td>14</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Poland</td>
<td>14</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Germany</td>
<td>12</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Italy</td>
<td>11</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Belgium</td>
<td>7</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Netherlands</td>
<td>7</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Egypt</td>
<td>5</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>Ukraine</td>
<td>5</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>Portugal</td>
<td>3</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>China</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>England</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>India</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Brazil</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Turkey</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Setting this threshold, all Turkish and Indian institutions were excluded. After this filter, the UM5S's potential partners are listed in Table 4. Neither publications nor citations of the Chinese Academy of Science is reported, because it is more a funding and policy-making body than a research institution, even if its output is almost 120,940 publications during the period of study. Similar to the institutions of classes 1 and 2, all the institutions of class 3 were found to have a higher impact on the scientific community worldwide (between almost 4 and 6 citations per publication) compared to that of the UM5S (1.82).

Table 4: List of the UM5S's knowledge citing institutions from emerging countries

<table>
<thead>
<tr>
<th>Institution</th>
<th>Nb. Publications (p)</th>
<th>Nb. Citations (c)</th>
<th>Impact (c/p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tongji University (China)</td>
<td>12,217</td>
<td>52,242</td>
<td>4.28</td>
</tr>
<tr>
<td>Sichuan University (China)</td>
<td>16,316</td>
<td>95,782</td>
<td>5.87</td>
</tr>
<tr>
<td>Shanghai Jiao Tong University (China)</td>
<td>27,993</td>
<td>123,550</td>
<td>4.41</td>
</tr>
<tr>
<td>Universidade De Sao Paulo (Brazil)</td>
<td>42,022</td>
<td>230,347</td>
<td>5.48</td>
</tr>
<tr>
<td>Universidade Federal Do Ceara (Brazil)</td>
<td>3,872</td>
<td>14,678</td>
<td>3.79</td>
</tr>
</tbody>
</table>

At the end of the method's stream, the institutions of the three classes are listed in Table 5, each requiring a specific action but within the same scientific collaboration policy.
Table 5: List of the institutions for collaboration with UM5S according to the three classes.

<table>
<thead>
<tr>
<th>Class 1: Institutions having agreements and co-publications</th>
<th>Class 2: Institutions having co-publications without agreements</th>
<th>Class 3: Institutions citing the university’s knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mohammed V-Agdal</td>
<td>University Hassan II</td>
<td>Tongji University</td>
</tr>
<tr>
<td>Ibn Tofail university</td>
<td>Antonoma Madrid</td>
<td>Sichuan University</td>
</tr>
<tr>
<td>CNRS</td>
<td>King Saud University</td>
<td>Shanghai Jiao Tong University</td>
</tr>
<tr>
<td>Zaragoza university</td>
<td>University Mohammed I</td>
<td>Universidade De Sao Paulo</td>
</tr>
<tr>
<td>Bordeaux university</td>
<td>University Czestochowa</td>
<td>Universidade Federal Do Ceara</td>
</tr>
<tr>
<td>Montreal university</td>
<td>University Lodz</td>
<td></td>
</tr>
<tr>
<td>Laval university</td>
<td>Kyoto University</td>
<td></td>
</tr>
<tr>
<td>Cadi Ayyad university</td>
<td>Hop Europeen Georges Pompidou</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nancy 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MC Erasmus (Rotterdam)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>University Valencia</td>
<td></td>
</tr>
<tr>
<td></td>
<td>University Rouen</td>
<td></td>
</tr>
<tr>
<td></td>
<td>University Liege</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CHU Tours</td>
<td></td>
</tr>
<tr>
<td></td>
<td>University Sidi Med Ben Abdallah</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Paris 7</td>
<td></td>
</tr>
</tbody>
</table>

4. Conclusion

A method is suggested for assessing a university scientific collaboration policy. Starting from data on agreements, co-publications, Research Fronts and citations, scientometric indicators were built along with science mapping in a comprehensive and organized method to inform about the coherence between co-publications as research collaboration outputs and collaboration agreements as a ‘legal’ frame input. The method stands to provide valuable inputs for policy assessment and design by identifying three distinctive institution classes for collaboration. The first class comprises institutions for which there were both co-publications and collaboration agreements; the university has to intensify its collaboration with this class.

The second class comprises institutions with significant co-publications and higher impact but no collaboration agreement; the university should consider setting up a formal ‘bottom-up’ frame of collaboration to better enhance existing individual collaborating activities. The third class comprises institutions highly citing university knowledge and having higher impact, thus providing potential collaborative partners.

Acknowledgement

The author would like to greatly thank R. Bouabid for his kind help.
References


Bouabid H. (2014), Science and technology metrics for research policy evaluation: some insights from a Moroccan experience, Scientometrics, 101, 899-915.


Choi S. (2012), Core-periphery, new clusters, or rising stars?: international scientific collaboration among ‘advanced’ countries in the era of globalization, Scientometrics, 90, 25-41.

Defazio D., Lockett A., Wright M. (2009), Funding incentives, collaborative dynamics and scientific productivity: Evidence from the EU framework program, Research Policy 38 (2), 293-305.


Lee S. & Bozeman B., (2005), The impact of research collaboration on scientific productivity, Social Studies of Science, 35(1), 673-702.
Levitt J. M., Mike Thelwell M. (2010), Does the higher citation of collaborative research differ from region to region? A case study of Economics, Scientometrics, 85 (1), 171-183.
Revealing existing and potential partnerships: affinities and asymmetries in international collaboration and mobility

Zaida Chinchilla-Rodríguez1, Yi Bu2, Nicolás Robinson-García3, Rodrigo Costas4, and Cassidy R. Sugimoto2, 4

1 zaida.chinchilla@csic.es
Consejo Superior de Investigaciones Científicas (CSIC) (IPP), SCImago Group (Spain)

2 buyi@iu.edu sugimoto@indiana.edu
Indiana University Bloomington (USA)

3 elrobinster@gmail.com
Universitat Politècnica de València (INGENIO-CSIC-UPV), Valencia (Spain)

4 rcostas@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (Netherlands)

Abstract
This study provides a preliminary analysis of the international profiles in collaboration and mobility using the seven countries indicated in the US executive order of January 27, 2017. The objective of this research is to analyze the flow of knowledge between countries and the relative importance of specific countries in order to inform evidence-based science policy. The work serves as a proof-of-concept of the utility of asymmetry and affinity indexes for collaboration and mobility. Comparative analyses of these indicators can be useful for informing immigration policies and motivating collaboration and mobility initiatives. Our analysis reinforces many of the established understandings of collaboration and mobility relationships—emphasizing the importance of geographic and cultural similarities. The analysis also explores the varied lenses on the importance of particular countries, when viewed from egocentric and relational perspectives. Our analysis suggests that comparisons of collaboration and mobility from an affinity perspective can identify gaps for mobility initiatives, given established scientific relationships. This approach can inform international immigration policies, but can also serve to identify potential partnerships at other levels of analysis (e.g., institutional, sectoral, or by state/province).

Conference Topic
Science communication; Science policy

Introduction and Background
Immigration is heralded as the key contemporary policy issue (Duncan, 2016; Bildt, 2017). Across the globe, political leaders are proposing and implementing nationalistic policies that restrict global mobility. One notable example is the executive order signed by United States President Trump on January 27, 2017 temporarily suspending entry of individuals from seven countries (i.e., Iran, Iraq, Libya, Somalia, Sudan, Syria, and Yemen) and placing restrictions on visa renewals for an additional 38 countries. Scientists home and abroad decried the executive order as impeding science, providing anecdotes of students, postdocs, and researchers who were trapped either home or abroad and unable to continue in their scientific activities (Morello & Reardon, 2017). However, there was little research on the scientific relationships among these countries and the implications of the policies for the scientific community.

Previous analysis showed the exponential growth of scientific connections at the global level over the last two decades (Adams, 2013; Wagner et al., 2015), primarily measured through collaboration (or co-authorship) (Newman, 2001; Barabási et al., 2002; Glanzel & Schubert,
2004; Leydesdorff & Wagner, 2008; Perianes-Rodriguez et al., 2009; Chinchilla-Rodriguez, et al, 2010; Gazni, et al., 2012; Wagner, Park, & Leydesdorff, 2015). Less understood is how scientists create relationship among countries through mobility. The use of bibliometric methods, in particular, has received scant attention, with a few notable exceptions (Moed & Halevi, 2014; Sugimoto, Robinson-Garcia, & Costas, 2016). However, there are traces of international relationships embedded in publication data: scholars co-author publications with scholars from other countries, they co-affiliate in multiple countries, and they change countries across their publication career. This data, however, has not been effectively harnessed to inform policy.

We provide here a proof-of-concept analysis, demonstrating the utility of bibliometric data to inform our understanding of the scientific relationship between countries, by using the countries indicated in the so-called “immigration ban” of January 2017. We present an approach that measures proportional production and probabilistic affinities for both collaboration and mobility, to describe the relationship in the context of the global scientific system. Furthermore, we compare these two indicators of international connectivity—collaboration and mobility—to provide a more refined understanding of the role of certain countries in the global knowledge network.

The vast differences in the volume and research capacities of different countries is a topic of concern for constructing valid indicators of international collaboration. In general, the more countries differ in scientific size, the larger the difference in relative terms (Luukkonen, Tijssen, Persson, & Silvertsen, 1993). Furthermore, the existence of collaboration and mobility between two countries implies reciprocity, but may not be symmetric. This is well-understood in network science: the degree of reciprocity is determined not by the definition of the link but by the extent to which two nodes report the same relationships with one another and two nodes exhibit proportional relationships based on overall performance of each one of all nodes in the network (Tichy, Tushman, & Fombrun, 1979). Several indices exist to measure network asymmetries (Luukkonen et al., 1993; van Eck & Waltman, 2009). Among them, two particular indices—Affinity Index (AFI) and Probabilistic Affinity Index (PAI)—have been used to measure the relative strength of networking science. The first one is size-dependent and the latter is size-independent. AFI allows for the calibration of relative importance and asymmetries between countries, measuring the amount of collaborative papers published jointly and the total number of international collaborations of each country and also, the number of active authors and the number of mobile authors. In a similar way, taking into account all the potential partners in the network and by removing the size dependency, PAI reveals the affinities that are largely dependent upon different drivers of scientific collaboration and mobility. In particular, the PAI is highly sensitive and reveals the limit of the scientific sizes as a predictor of relationships. Previous studies have employed AFI and PAI (e.g., Zitt et al., 2000; Finardi & Buratti, 2016) to analyze collaboration, but no previous study has, to our knowledge, applied these measures to analyze the connections constructed through mobility.

We examine, for each of the seven countries for which visas were temporarily suspended as well as the United States, four dimensions: (a) asymmetry of collaboration, (b) asymmetry of mobility, (c) affinity for collaboration, and (d) affinity for mobility. We provide collaboration and mobility in parallel to highlight the different insights that can be gained by examining these indicators in concert, rather than independently. Our objective is to provide bibliometric tools to bring scientific evidence to arguments of anecdote in this time of heightened concern over globalization.

Data and Methods

By using affiliation data from scientific publications, it is possible to track the trajectory of individual scientists and analyze collaboration and mobility at the level of research groups,
subject areas, and countries (Moed & Halevi, 2014; Sugimoto, Robinson-Garcia, & Costas, 2016). In this study, we use co-authorship as a measure to examine international collaboration and the number of affiliations of scholars with different countries as a proxy of international mobility in a sample composed of the seven countries involved in the immigration ban--Iran, Iraq, Libya, Somalia, Sudan, Syria and Yemen--and the United States.

The data is drawn from a curated version of the Web of Science housed at CWTS (Leiden University). Author data from 2008 to the present has been disambiguated using an author disambiguation algorithm developed by Caron and van Eck (2014), wherein individuals are matched with distinct publication records based on similarities in co-authorship networks and affiliation information. The use of a disambiguation algorithm is necessary to identify mobility at the individual level. We therefore restrict our data to the period 2008-2015, to utilize this disambiguated data (14,097,939 publications). We further refine our data only to those links which contain at least one of the countries of interest in our sample and include only those individuals who have at least 2 publications (n=3,521,797). These nation-to-nation links were associated with 8,168,640 publications (2008-2015) across 216 countries. For collaboration, we considered all authors listed on the papers and their connections to countries. Reprint affiliations have been included for both collaboration and mobility analysis and all document types were included.

Country names were cleaned and standardized (e.g., merging Timor-Leste and East Timor, Congo and Republic of Congo as well as Zaire and the Democratic Republic of Congo). The cleaning resulted in 213 countries. We then enriched this data by adding additional classifications. Every country was assigned to one of eight geographical regions: North America, Latin America and the Caribbean, Africa, Asia, the Middle East, the Pacific Region, Eastern Europe, and Western Europe. Income level category, as defined by the World Bank (2016) was used as a proxy for economic capacity. The Scientific Technological and Capacity Index (Wagner et al., 2001) was used as an indicator of scientific capacity. This index categorizes countries into four groups of unequal sizes: advanced (n=22), proficient (n=24), developing (n=22), and lagging (n=80). In the present study, an additional group (“other”) is added to account for the 66 countries that were not placed in the Wagner et al. (2001) classification.

Full-counting is used to attribute the credit to authors for collaboration and mobility. Thus, the sum of publications and authors exceeds the total number of publications. From these data, we calculate the following indicators:

- total number of publications per country (# papers);
- position in the ranking of the number of publications at the international level (Rank # papers);
- number of publications with international collaboration (#papers in international collaboration);
- proportion of papers with foreign partners (% international collaboration);
- number of countries involved in international collaboration (# collaborators countries);
- percentage of countries reached in international collaboration with respect the total number of potential collaborators (% collaborators countries);
- number of authors per country (# authors),
- number of international mobile authors per country (# authors in international mobility);
- percentage mobile authors with respect the total authors (% mobile authors);
- number of countries with mobility connections (#mobility countries),
- percentage of countries reached in international mobility with respect the total number of potential countries (% mobility countries);
- ratio of the proportion of collaborators countries among all potential collaborators;
and the proportion of the countries with mobility connections among all potential countries (Ratio collab-mobil). For example, Yemen is collaborating with 109 out of 212 countries reaching around 50% of potential collaborators, at the same time is reaching around the 12% of potential mobile connections. That means that Yemen multiplies by four the number of countries in collaboration respect those with mobility (Table 1).

To measure the asymmetry in collaboration and mobility patterns, we use a pair of inclusion indexes, $AFI(i,j)$, and counterpart, $AFI(j,i)$, with one-way normalization. AFI is a measure of the links between a given country (i) with another country (j), compared to the total links country (j) with the entire world during the same period. $AFI(i,j)$ is calculated as:

$$AFI(i,j) = \frac{n(i,j)}{\sum_j n(i,j)}$$

where $n(i,j)$ is the volume of links between countries i and j. This index is size-dependent given that the preference for partner j in $AFI(i,j)$ is influenced by the global size of j (conversely, preference is dependent upon the size of i for $AFI(j,i)$). This index highlights the important partners in terms of quantity and demonstrates the asymmetry in partnership.

To normalize by the size of both countries, we employ the Probabilistic Affinity Index (PAI), widely used in science policy to demonstrate the degree to which proximity (both material and immaterial) contributes to scientific relationships among the countries (Zitt et al., 2000). The index is capable of demonstrating strong relationships with small countries and presents the strength of relationships in their global context. For PAI, we refer to Zitt et al. (2000)’s algorithm and apply it to calculate the relationships between countries for both collaboration and mobility:

$$PAI(i,j) = \frac{[n(\ldots)n(i,j)/n(i)n(j)]^2 - 1}{[n(\ldots)n(i,j)/n(i)n(j)]^2 + 1}$$

where $n(i) = \sum_j n(i,j)$ and $n(\ldots) = \sum_{i\neq j} n(i,j)$. PAI is normalized into [0,1] in order to make it comparable with AFI.

Each indicator yields a different lens on the international portfolio of a given country. The affinity index reveals asymmetries, demonstrating that collaborations with country i might constitute a large portion of the total collaborations of country i, but a small proportion of the collaborations of country j. It simultaneously provides an indicator of the main collaborators in the global network. PAI, on the other hand, highlights the strengths of countries that might not have high output, but have disproportionately strong connections in the global environment. We present, for both collaboration and mobility, the 50 main and preferred partners for the countries in our sample, according to the AFI and PAI.

**Results**

**Overview of collaboration and mobility**

The seven countries differ in both economic and scientific dimensions. Iran, Iraq, and Libya are economically considered as upper-level income countries, whereas Sudan, Syria, and Yemen are in the lower-middle income bracket for countries. Somalia is in the lowest income level. Somalia is also distinct in terms of the profile of scientific output and the size of the scientific workforce—having the lowest of both for the seven (Table 1). Iran is the most prolific country in the set—contributing 1.8% of the share of world publications, placing it in 17th place worldwide in terms of scientific output. The next most productive country in the set is Iraq, which is ranked 82nd globally in terms of production and produces 0.05% of the world output.
The general international profile of the countries also differs—nearly all publications from Somalia are authored with an international partner, whereas only a fifth of publications from Iran are the result of international collaboration. Iran and Sudan have the widest collaboration portfolios in terms of number of international partners—reaching 76 and 75% (respectively) of potential collaborative partners in this dataset. Somalia has the fewest, reaching less than 20% of potential partnerships, likely as a result of the low output.

Mobility provides a different lens on international partnerships. For example, Iraq has one of the lowest rates of collaboration, yet nearly half of the researchers in Iraq have some degree of mobility and reaches nearly a quarter of all potential partners. Iran is the only country that remains stable across the two lists—appearing as both the least collaborative and the country with the lowest percentage of mobile researchers. However, even with the lower degree of mobility, it still reaches nearly 40% of potential mobility countries, second in this set only to the United States (at 88%). These results suggest that collaboration and mobility indicators present different images of the international profiles of countries. A comparison of the indicators suggests that mobility is a more selective indicator: in all countries, the ties between collaborative countries is higher than the ties established through mobility. In a way, mobility is a costlier (particularly from a human, social and economic point of view) event than collaboration, thus explaining the prominence of collaboration over mobility.

### Table 1. Percentage of international collaboration, collaborators countries, authors in mobility and mobility countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>Code</th>
<th>Rank #</th>
<th>papers in international collaboration</th>
<th>% international collaboration</th>
<th># collaborators</th>
<th>% collaborators countries</th>
<th># authors</th>
<th>% authors in mobility</th>
<th># mobility authors</th>
<th>% mobility countries</th>
<th># mobility countries</th>
<th>ratiocollab-mobil</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>1</td>
<td>2292220</td>
<td>753271</td>
<td>53.86</td>
<td>77697</td>
<td>9.8</td>
<td>188</td>
<td>88.68</td>
<td>1.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRAN</td>
<td>17</td>
<td>146883</td>
<td>31695</td>
<td>21.58</td>
<td>4752</td>
<td>8.3</td>
<td>85</td>
<td>40.09</td>
<td>1.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRAQ</td>
<td>82</td>
<td>3759</td>
<td>2508</td>
<td>66.72</td>
<td>662</td>
<td>47.7</td>
<td>49</td>
<td>23.11</td>
<td>2.28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUDAN</td>
<td>99</td>
<td>1853</td>
<td>1487</td>
<td>80.25</td>
<td>288</td>
<td>39.2</td>
<td>46</td>
<td>21.70</td>
<td>3.44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYRIA</td>
<td>103</td>
<td>1575</td>
<td>1122</td>
<td>71.24</td>
<td>195</td>
<td>36.0</td>
<td>40</td>
<td>18.87</td>
<td>3.29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YEMEN</td>
<td>117</td>
<td>1075</td>
<td>948</td>
<td>88.19</td>
<td>152</td>
<td>49.8</td>
<td>27</td>
<td>12.74</td>
<td>3.96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIBYA</td>
<td>119</td>
<td>1002</td>
<td>841</td>
<td>83.93</td>
<td>217</td>
<td>49.5</td>
<td>33</td>
<td>15.57</td>
<td>3.54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOMALIA</td>
<td>194</td>
<td>40</td>
<td>39</td>
<td>97.50</td>
<td>3</td>
<td>15.0</td>
<td>2</td>
<td>0.94</td>
<td>19.14</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The dominant partners for both collaboration and mobility are scientifically “advanced” countries (Table 2). Yemen is an exception, collaborating predominately with “lagging” countries. All countries have a demonstrable share of connections with lagging countries, when compared to “developing” and “proficient” countries showing homophily in their international relationships.

### Table 2. Percentage of collaboration and mobility with countries by S&T Capacity Index

<table>
<thead>
<tr>
<th>S&amp;T Index</th>
<th>Mobility Collab</th>
<th>Mobility Collab</th>
<th>Mobility Collab</th>
<th>Mobility Collab</th>
<th>Mobility Collab</th>
<th>Mobility Collab</th>
<th>Mobility Collab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced</td>
<td>42.94</td>
<td>42.94</td>
<td>42.94</td>
<td>42.94</td>
<td>42.94</td>
<td>42.94</td>
<td>42.94</td>
</tr>
<tr>
<td>Proficient</td>
<td>13.95</td>
<td>13.95</td>
<td>13.95</td>
<td>13.95</td>
<td>13.95</td>
<td>13.95</td>
<td>13.95</td>
</tr>
<tr>
<td>Lagging</td>
<td>29.41</td>
<td>29.41</td>
<td>29.41</td>
<td>29.41</td>
<td>29.41</td>
<td>29.41</td>
<td>29.41</td>
</tr>
<tr>
<td>Others</td>
<td>2.35</td>
<td>2.35</td>
<td>2.35</td>
<td>2.35</td>
<td>2.35</td>
<td>2.35</td>
<td>2.35</td>
</tr>
</tbody>
</table>

Western Europe is also a dominant partner, but with some exceptions (Table 3). For example, the plurality of Iraq’s mobility and collaboration partners come from the Middle East. Somalia’s mobility is limited to Africa and Asia; and Sudan and Yemen have mobility largely with Asian countries.

### Table 3. Percentage of collaboration and mobility with countries by geographic region

<table>
<thead>
<tr>
<th>Regions</th>
<th>Mobility Collab</th>
<th>Mobility Collab</th>
<th>Mobility Collab</th>
<th>Mobility Collab</th>
<th>Mobility Collab</th>
<th>Mobility Collab</th>
<th>Mobility Collab</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Africa</td>
<td>9.72</td>
<td>7.85</td>
<td>4.31</td>
<td>8.57</td>
<td>7.78</td>
<td>23.44</td>
<td>22.44</td>
</tr>
<tr>
<td>% Asia</td>
<td>18.87</td>
<td>17.63</td>
<td>44.81</td>
<td>31.52</td>
<td>12.94</td>
<td>14.66</td>
<td>33.33</td>
</tr>
<tr>
<td>% Eastern Europe</td>
<td>2.85</td>
<td>13.46</td>
<td>2.84</td>
<td>5.21</td>
<td>1.57</td>
<td>4.74</td>
<td>0.30</td>
</tr>
<tr>
<td>% Latin America&amp; Caribbean</td>
<td>0.81</td>
<td>4.07</td>
<td>3.04</td>
<td>0.78</td>
<td>3.45</td>
<td>4.69</td>
<td>4.10</td>
</tr>
<tr>
<td>% Middle East</td>
<td>2.64</td>
<td>6.90</td>
<td>8.96</td>
<td>16.89</td>
<td>27.06</td>
<td>23.81</td>
<td>7.81</td>
</tr>
<tr>
<td>% Northern America</td>
<td>30.76</td>
<td>16.66</td>
<td>9.39</td>
<td>8.82</td>
<td>7.45</td>
<td>6.84</td>
<td>10.16</td>
</tr>
<tr>
<td>% Pacific Region</td>
<td>7.34</td>
<td>5.10</td>
<td>6.26</td>
<td>3.71</td>
<td>4.71</td>
<td>3.39</td>
<td>3.13</td>
</tr>
<tr>
<td>% Western Europe</td>
<td>36.01</td>
<td>34.33</td>
<td>27.03</td>
<td>26.50</td>
<td>44.71</td>
<td>34.54</td>
<td>42.19</td>
</tr>
</tbody>
</table>
Country-specific analysis

**Iran.** High producing and advanced countries, such as the United States and Canada, lead in terms of the top connections for both collaboration and mobility, when considering the AFI for Iran (Figure 1A). For example, Iran publishes more than 22% of internationally-collaborative publications with the US and 20% of Iranian researchers (those for whom their first publication was in Iran) have been affiliated with the US. However, Iran also has a strong connection with Malaysia—a scientifically lagging country with upper middle income. Among the countries with the strongest ties, collaboration is stronger with the US, Germany, and the United Kingdom than mobility, whereas the inverse is true with Canada and Malaysia. Overall, using AFI as the measurement, Iran tends to have stronger relationships in terms of collaboration than mobility.

PAI presents a very different portfolio in terms of international connections. The strongest partners using this index include Azerbaijan, Malaysia, Armenia, and Iraq—reinforcing the importance of geographical, linguistic, and cultural ties. Canada also remains highly ranked according to this index. The United States, however, drops considerably in the ranking and the United Kingdom show a negative index in terms of mobility. As with AFI, the connection with the US on the PAI is stronger with collaboration than mobility. This preference is invariable across the 50 preferred partners for Iraq. Several countries have a very high probability index for collaboration, but negative ranks for mobility. These include several Latin American (i.e., Brazil, Mexico, Chile) and Middle Eastern (i.e., Pakistan, Saudi Arabia, Egyptian) countries. This may suggest potential avenues for mobility programs.

**Iraq.** Iraq collaborates with 115 countries and has researchers with mobility connections with 50, including Palau and Luxembourg, with whom there are no collaborations in common (Figure 1B). As measured with AFI, Malaysia captures the largest share of collaboration and mobility relations with Iraq, followed by the UK, the US, and China. Countries such as the US, Italy, Jordan, and Saudi Arabia have demonstrably higher AFI indexes in collaboration than mobility; however, for the most part there is an evenness in the collaboration and mobility patterns for Iraq according to AFI.

Taking into consideration all potential partners with PAI, the strongest relationships are with Malaysia, Kuwait, Brunei, Libya, Yemen, Afghanistan, Syria, Oman and Jordan (PAI 0.99 and 0.95 in collaboration and mobility). Among the most developed countries only the United Kingdom, Australia, and Sweden are preferred partners (with PAI values of 0.75 and 0.7 in mobility and 0.93 and 0.87 in collaboration). In contrast, Iraq shows low affinities with China, Germany, and the USA. This suggests that some scientifically developed countries located in...
Europe, the Pacific region, and Asia are acting as bridges of enhancing the connectivity of Iraqi science. As with Iran, the comparison between collaboration and mobility using PAI reveals gaps in partnerships. For example, Iraq has strong affinities for collaboration with Saudi Arabia and Israel (0.99 and 0.93, respectively), but no mobility. There are, of course, political explanations for these discrepancies, but the measurement provides a neutral way to identify communities that have demonstrated scientific affinities (i.e., through collaboration) which could be explored in further policy initiatives.

**Sudan.** As seen in Figure 2, Sudan has mobility with 46 countries (39%) and collaborates with 161 (74.5%), a similar figure to the one observed in Iran despite the difference in number of publications. In terms of size, Sudan’s main partners in mobility are Saudi Arabia, China, Malaysia, Germany, South Africa, the United Kingdom and the USA (5.5%); while the collaborative partners are different—the US is the top collaborative partner followed by Saudi Arabia, the United Kingdom, Germany, and China. Similar to Iran, collaboration tends to be much higher in all cases than mobility, as measured by AFI.

When correcting for size, the strongest partners are Eritrea, Bahrain, Brunei (in mobility, but not collaboration), Saudi Arabia, Malaysia, South Africa, Uganda, Yemen, Cote Ivoire, Ethiopia, Zambia, Oman, and Kenya. In the case of Sudan, neighbors and language matters in the strength of the relationships with the exception of Malaysia. The United States has a fairly high collaboration PAI, but is low in terms of mobility, similar to other advanced countries, such as France, Switzerland, and Canada. This may indicate avenues for potential collaboration between these advanced countries and Sudan, a lagging country.

![Figure 2. Main partners of Sudan (left A) and Syria (right B) in mobility and in collaboration according to AFI (top) and preferred partners in PAI (bottom).](image)

**Syria.** Nearly three-quarters of Syria’s output (71.2%) is the result of international collaboration, written in collaboration with 134 countries. More than a third of Syrian authors have mobility, linking Syria with 40 countries. As with Sudan, Syria does not have a parallel relationship between collaboration and mobility—the US, for instance, is highest in collaboration by AFI, but is ranked fourth by mobility. France, Germany, and the UK are strong partners on both dimensions. Across the top 50 countries, Syria has a pronounced emphasis on collaboration over mobility.

PAI again demonstrates the importance of socio-cultural factors in establishing collaborative ties. Syria’s preferred partners in this index include Yemen, Uzbekistan, Jordan, Afghanistan, Morocco, Lebanon, Ethiopia, Iraq, Sri Lanka, United Arab Emirates, Saudi Arabia, and Qatar. A second preferential zone is formed by European countries—including France, Turkey, Germany, Austria, the United Kingdom, Hungary, and Portugal. The US is not a preferred partner, demonstrating low degrees of mobility, despite fairly high levels of
collaborative affinity. Syria has several countries in its profile with extreme differences in affinity between collaboration and mobility, including Oman, Tunisia, Algeria, Pakistan, Serbia, Argentina, and Israel.

**Yemen.** Yemen collaborates with 109 countries and has mobility connections with 27 countries. (these are not entirely overlapping units, as there is mobility, but no collaborative publications with Madagascar). The strongest partners in terms of size include Malaysia, Saudi Arabia, and Egypt. By the AFI, the US is ranked fourth in terms of collaboration and sixth by mobility. Yemen demonstrates higher degree of collaboration than mobility by AFI.

There are stronger similarities between preferred partners by AFI and PAI. Using PAI, the strongest ties are with Syria, Malaysia, Saudi Arabia, Egypt, Qatar, and Sudan. When considering the entire global context, Yemen has several countries for whom there are less than expected values for both mobility and collaboration, including South Korea, Sweden, Canada, Japan, Australia, and Spain.

**Libya.** Libya has collaborative partnerships with 119 countries and mobility with 33 countries (including Lesotho, Kyrgyzstan, and the West Indies with which Libya has no collaborative papers). In gross volume, the main partners in collaboration are the United Kingdom, Egypt, France, India, Malaysia, and the USA (9.8% in collaboration and 2.3% in mobility). Collaboration is the dominant form of connection.

The preferred partners are Lebanon, Angola, Malaysia, Iraq, Serbia, Oman, Qatar, Finland, and Greece. The second zone of influence comprised the United Kingdom, United Arab Emirates, France, and India. The Unites States is not a preferred partner. Libya’s collaboration and mobility ties with many African countries are much lower than would be expected, given the global scientific network.

**Somalia.** Somalia has the lowest rate of publication across the seven countries, with collaborative ties with 40 countries and mobility ties with only two—Kenya and Malaysia. The US is the fourth most important partner in AFI, after Kenya, Malaysia, and the United Kingdom. Somalia is outperforming expected collaboration rates in all countries, but is underperforming expected values of mobility in all countries save Kenya and Malaysia.

**United States.** Given that our proof-of-concept study focuses on the relationship of the US with the seven countries in the executive order, it is necessary to also take an egocentric analysis of the US for context. The US has one of the lowest levels of international collaboration due to size and scientific capacity, but serves as the dominant collaboration for nearly all countries across the world. We provide both the top 50 connections, but also the countries in examination...
(the fact that these do not overlap suggests the comparatively weaker relationship between these countries from the perspective of the US). According to AFI, the strongest partners include (in order of importance) China, the UK, Germany, Canada, France, Italy, Japan, South Korea, and Switzerland. Collaboration rates are higher than mobility rates for all countries except for China and Iran, countries with the exception of China and Iran where the percentages are slightly higher (China: 18% for collaboration and 19% mobility; Iran 1% for both collaboration and mobility).

**Figure 4.** Main partners of USA in mobility and in collaboration according to AFI (top) and preferred partners in PAI (bottom).

When size dependence is removed, Israel is the main collaborative partner with the US, followed by China, South Korea, Tukey, Taiwan, Peru, Lebanon, Uganda, and Canada. Unlike the other countries studied, Malaysia is not a preferred partner. In the global market, the US is outperforming in terms of collaboration across all the countries. However, it has lower than expected values in terms of mobility with Pakistan, Belgium, Saudi Arabia, and Malaysia (within the top 50 counties) and with six of the seven countries (Iran is the exception).

**Discussion and Further Research**

This work presents a case study of the application of asymmetry and affinity measures to better understand the implications of migration policies. Using the countries suspended in the 2017 US executive order, we present an egocentric view of each country involved in order to analyze the strength and reciprocity of the relationships of each of the various countries with the US and with other preferred partners. Although affinity indexes have long been used to analyze international collaboration, this is the first analysis to explore the insights gained when mobility and collaboration are analyzed in parallel. For the case under study, we demonstrated an asymmetrical relationship between each of these countries and the US—the US was a dominant country in terms of collaboration and mobility for the seven countries, though they represented a small fraction of the share of US scientific relationships. Of the seven, Iran was the most important scientific country for the US, with higher than expected amounts of collaboration and mobility. Across the countries, Malaysia was a consistently preferred partner—a particularly striking relationship given the low affinity between the US and Malaysia.

Internationalization of science affects the research performance of countries and in certain degree depends on the attractiveness of a partner in the global network. The country with which the scientific relationship is established is important for indicator construction,
given unequal magnitude of contributions between partners. The results of this study demonstrate that, despite the low volume of international publications and mobile researchers for many countries, the number of countries reached is relatively high. This internationalization presents policy challenges and opportunities, particularly for developing and lagging countries that are simultaneously confronted with critical internal conditions that often isolate them from countries with high scientific capacity. Despite these barriers, international collaboration seems to respond to the dynamics created by the self-interests of individual scientists rather than to other structural, institutional, or policy-related factors (Leydesdorff & Wagner, 2008). This is evident in the degree of internationalization of the countries in our sample.

Previous studies have revealed different drivers underlying the formation of collaborations networks such as geographical and cultural proximity, thematic similarity, regional partnership, as well as the reputation and attractiveness of the receiving partner (Zitt, Bassecoulard, & Okubo, 2000; Archambault, Beauchesne, Côté, & Roberge, 2011; Adams et al. 2014; Finardi & Buratti, 2016). Our results confirmed these pre-existing notions: AFI demonstrated the asymmetries and preferential attachments to countries with high reputation and scientific capacity. PAI reinforced the importance of geographic and linguistic proximity. Large countries, like the US, have extremely high breadth across the scientific network—serving as a collaborative partner with all and a mobile partner for many. Smaller countries, in turn, rely upon international collaboration for a large share of their output. However, there is a distinct lack of reciprocity in terms of mobility. Whereas there are established scientific similarities, on the basis of patterns of collaboration, this does not necessarily lead to increased mobility. Mobility, therefore, is a more selective indicator; while collaborations can be established without colocation, mobility—even in the case of co-affiliation—(i.e., researchers with a double appointment with two or more international institutions but keeping ties with their country of origin) often requires some degree of travel and human interaction, having a higher social, human, and economic cost. Mobility is thus likely to be heavily influenced by political environments. Given that the countries in our sample have experienced embargos, invasions, civil wars, and revolutions in the course of the time period under investigation, these are likely to have influenced mobility rates (Moed, 2016).

Scientific relationships are highly resource-dependent (Pouris & Ho, 2014). However, our analysis has demonstrated the potential utility of combining collaboration and mobility indicators to identify preferred partners, inform policies, and identify potential areas for establishing mobility programs. This method could be extended to multiple levels of analysis, for example, to the institutional level, to examine intra- and inter-national collaboration networks. Further research should investigate the degree to which mobility and collaboration initiatives are successful in facilitating relationships between partners, particularly those that are not already established as preferred or dominant partners, and how to enhance the attractiveness of countries in the global scientific system.

**Limitations and further analysis**

This study requires further analysis in order to overcome some of the limitations and respond to other important questions related to the capacities and influences of countries in science. The data examines only those authors whose first publication occurred in or after 2008. This provides only limited diachronic analysis. More data and analysis are necessary to further inform a longitudinal analysis of collaboration and migration patterns, especially due to the inflationary effect on the traditional measurements of collaboration based on affiliations, that creates some overlaps when we are comparing collaboration and mobility. There is a major general limitation of collaboration analysis based on author affiliations and further analysis should be done to minimize this effect. In further research, we intend to complement our analysis with a time component, allowing us to analyze some key points such as authors’
choices regarding institution address selection from publication data. This will facilitate the analysis of causal relationships, examining the relationship between collaboration and migration as well as the effects of policies and political action (Hottenrott and Lawson, 2017). We also plan to analyze positions occupied in the bylines of co-authorship, the impact of publications as a result of these relationships, the institutional reputation of destinations, and the degree to which topic changes occur as a result of these interactions. At the methodological level, approaches with different counting methods (Perianes-Rodríguez et al., 2016) and scale-adjusted metrics will be explored in order to assure an accurate comparison of relational capacities of countries with different sizes and capacities (Archambault et al., 2011; Finardi & Buratti 2016). There are also several other elements, such as the thematic specialization of science, which should be taken into account when analyzing collaborative preferences between countries (Glanzel 2000; Radosveic & Yoruk 2014).

Acknowledgments

Financial support from Mobility Program ‘Salvador de Madariaga 2016’ and State Programme of Research, Development and Innovation oriented to the Challenges of the Society (CSO2014-57770-R) funded by the Ministry of Economy and Competitiveness of Spain and the Science of Science Innovation and Policy program of the National Science Foundation in the United States (NSF #1561299).

References

Duncan, P. Immigration is key issue with EU referendum voters, according to Google. The Guardian, Retrieved April 7, 2017 from: https://www.theguardian.com/politics/2016/jun/15/eu-referendum-immigration-key-issue-with-voters-google


The structure and evolution of scientific collaboration from the perspective of symbiosis

Junwan Liu¹, Kaiyue Ding², Xiaomin Zheng³, Feifei Wang⁴
¹liunjunwan@bjut.edu.cn, ²2500976798@qq.com, ³292908015@qq.com, ⁴feifeiwang@bjut.edu.cn
School of Economics and Management, Beijing University of Technology, 100124 Beijing (China)

Abstract
With the continuous deepening and development of scientific collaboration, relationships within the scientific collaboration network are becoming more and more complex, exhibiting symbiotic characteristics. From the viewpoint of symbiosis, we examine the typical teacher–student relationship using social network analysis. We study the network characteristics of this typical form of collaboration based on the resultant published research papers, considering the contribution of symbiosis in the scientific collaboration index to analyze the evolution of the symbiotic relationship between collaborative objects. The results indicate that the symbiotic relationship between teacher and student shows obvious variations from the “old leads young” model; young researchers draw support from the teacher’s resources to improve their own level of scientific research with continuous progress and development. Moreover, the collaborative symbiosis degree between teacher and student changes with time.

Conference Topic
Social network analysis; Studies on the level of individual scientists

Introduction
Scientific research collaboration has become a significant feature and an important driving force in today’s scientific development; scientific collaboration is conducive to the realization of knowledge complementarity, resource integration, and interdisciplinary research. Scientific collaboration exhibits new features as the relationship between the collaborative subjects continuously deepens and evolves. However, scientific collaboration has transformed from the early point-to-point collaboration and chained collaboration pattern to the complex collaboration network mode, which resembles biological symbiosis between the collaborative subjects. This has prompted us to develop a theoretical framework and model to describe and explore this new scientific collaboration. Currently, interest is mostly focused on the network characteristics of scientific collaboration, and network theory has been used to study the structure and evolution of scientific collaboration. However, network theory can only reveal the external characteristics of scientific collaboration and does not consider the intrinsic motivation for the collaboration itself. On the contrary, symbiosis theory provides us with the appropriate perspective. Modern symbiosis theory is no longer confined to biology and has been extended to the fields of economics and social sciences, where it has been used to explain enterprise symbiosis relationships, industry–university–research symbiosis networks, etc.
In the era of big science, scientific collaboration has become the main pattern of scientific research; thus, research on scientific collaboration has also emerged steadily. After the twenty-first century, new research has been emerged in the field of scientific collaboration. Social network techniques and bibliometric indicators have been applied in the field of scientific collaboration (Hou et al., 2013). In fact, Newman’s publication on this subject has aroused widespread concern, reaching 1790 citations as of April 27, 2017. Newman used the Medline database to investigate the structure of scientific collaboration networks in biological medicine, physics, and computer science. The results show that scientific collaboration exhibits high clustering properties; the co-author network presents a typical small-world phenomenon (Newman, 2001) (reference [16]). Newman regarded that when constructing a collaboration network, authors of joint publications have an associative relation; in fact, this association definition is widely used in the research of scientific collaboration network. Research on structure of scientific collaboration using bibliometric indicators published by Glanzel (2002) received continued attention. Afterwards, bibliometric indicators, such as the number of papers, average number of authors, number of collaborators, maximum size of the scientific collaboration team, and aggregation degree, have been widely applied in the research of scientific collaboration networks (Cavusoglu & Turker, 2014; Wang et al., 2015). Social network analysis has been proven to be very effective for analyzing and visualizing co-authorship networks (Huerta-Barrientos et al., 2014; Otte & Rousseau, 2002; White, 2003; Kretschmer & Aguillo, 2004). Subsequently, complex network analysis methods were used for investigating the structure of scientific collaboration networks. Ding (2011) combined topic modeling and path-finding algorithms to study the collaboration and citation tendencies of productive authors in the field of information retrieval.

Research on the structure of scientific collaboration networks is a frontier subject in the field of scientometrics. Egghe (2003) and Newman (2001) studied scientific weighted collaboration networks and indicated that the distribution of weights reflects the activity of academic exchanges in this field; specifically, they noted that active scientists tend to collaborate with or aid other active scientists. Qiu et al. (2014) utilized the frequency statistical method for the first time to construct a network of the collaborative age of scholars in the field of informatics. Homophily and preferential attachment are two common mechanisms in a network’s evolutionary process. For these mechanisms, Wang and Zhu (2014) observed a simultaneous change in the regulations in the evolutionary processes of a scientific collaboration network and revealed that there are random selection, restrictive selection, and active selection features of collaborative links in the early, middle, and final stages of the evolution of collaboration networks respectively. Chen et al. (2013) analyzed the evolution of co-authorship networks in the journal Scientometrics from the micro and the macro angle.

The concept of “symbiosis” was first proposed by a German biologist named Anton de Bary in 1879; this concept was associated with survival needs, where two or more organisms gradually form a symbiotic relationship of coexistence and coevolution.
based on a particular pattern of interdependence and interaction. Yuan (1998) referenced the concept of symbiosis and developed related theories. He constructed the framework of “symbiosis theory” in economic analysis, using symbiosis theory research to solve practical economic problems in China and get some new ideas, perspectives, and methods. In recent decades, symbiosis theory has been gradually extended to the field of Social Sciences and Management Science and has achieved initial success, resulting in some research hotspot such as “the industrial symbiosis network” (Wang et al., 2005; Albino et al., 2015; Domenech et al., 2011), “enterprise cluster symbiosis” (Wang et al., 2006), and “the industry–university–research symbiosis network” (Feng et al., 2013). In the background of highly integrated knowledge, using symbiosis theory to investigate the characteristics of academic collaboration is also an innovative and challenging work. Shi et al. (2013), from the perspective of symbiosis theory, using the methods of literature research, social network analysis, mathematical statistics, and quantitative analysis, analyzed the inter-provincial cooperation mechanism and the regional characteristics of sports scientific research in China. Smith (2011) performed a quantitative analysis of citations from Wikipedia articles to documents in institutional repositories and revealed a potential symbiotic relationship between Wikipedia and academic research in the institutional repository. However, other research have not been found for the measurement of the scientific cooperation based on the symbiotic theory. These previous findings on the structural characteristics and evolution trend of scientific collaboration networks obtained via social networking techniques are quite relevant to the scope of the present research. However, the symbiotic relationships within the framework of scientific collaboration have not been investigated. In scientific collaboration relationships, there is a complementarity between the partners regarding knowledge, expertise, and resources. Therefore, there is the possibility of “symbiosis”. It is of great theoretical and practical significance to reveal the structure and evolution of scientific collaboration networks from the perspective of symbiosis. In this study, we reveal the scientific symbiosis network structure and its evolutionary trends using the symbiosis degree index by analyzing the scientific collaboration symbiotic relationship between academicians of the Chinese Academy of Sciences and their collaborators.

Data and Methods

Data
The academicians of the Chinese Academy of Sciences represent the highest academic level of Chinese scientists, play an important role in the development of the country and its society, and enjoy high social prestige. In a preliminary work, we studied the characteristics of a scientific collaboration network between academicians and found that there was a long-term and stable partnership between academicians and between academicians and non-academicians (Liu et al., 2015). Especially, in terms of the teacher–student relationship, the collaboration between Li Shushen and Xia Jianbai is the most typical example. The scientific collaboration between the two academicians lasted more than 15 years, and they co-authored at least one paper every
year. Thus, we firmly believe that their scientific collaboration showed symbiotic characteristics.

The official website of the Chinese Academy of Sciences provides the basic information of the academicians, such as gender, date of birth, place of origin, time of election, work units, as well as their main fields of research experience and award-winning situations. We selected the Web of Science database as the paper search source database. Using the full names and the abbreviations of the English names of the academicians, We can get their SCI publications. After that, the search results were screened manually according to their institutions, research areas, and other information regarding their published works to acquire the exact number of all their publications.

In this study, we choose Li Shushen and Xia Jianbai to investigate the teacher-student collaboration pattern (shown in Table 1). They are academicians of the Information Technology and Science Department, have been working at the Institute of Semiconductors of the Chinese Academy of Sciences, and have maintained long-term and frequent collaborative relations. Xia Jianbai obtained his master’s degree from the Department of Physics of Peking University in 1965, and he was elected as an academician of the Chinese Academy of Sciences in 2001. Li Shushen received his doctor’s degree from the Institute of Semiconductors of the Chinese Academy of Sciences in 1996, and he was elected as an academician of the Chinese Academy of Sciences in 2011. They published their first co-authored paper in 1994, when Li Shushen was a Ph.D. student at the Institute of Semiconductors; thus, this paper affirms that their relationship was a collaboration between teacher and student. For the next 20 years, they maintained a collaborative relationship without interruption. Here, we define this cooperative relationship as the symbiosis relationship. So is there any significant change in the symbiosis degree of scientific collaboration between a teacher and student in long-term collaboration? In this paper, we perform an empirical analysis on the symbiotic relationship between teacher and student in the form of collaboration.

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Title</th>
<th>Time of being elected as an academician</th>
<th>Number of co-authored papers</th>
<th>Duration of collaboration (year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xia Jianbai</td>
<td>77</td>
<td>Academician</td>
<td>2001</td>
<td>90</td>
<td>20</td>
</tr>
<tr>
<td>Li Shushen</td>
<td>53</td>
<td>Academician</td>
<td>2011</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Measurement of symbiosis degree

Symbiosis is a self-organizing process, wherein mutual understanding and collaboration are established between the symbiotic units; moreover, there is complementarity and interdependence in terms of material resources, information, and energy. In biology, symbiosis is considered one of the fundamental mechanisms leading to species innovations that trigger species evolution. In sociology, this symbiotic relationship promotes economic innovation, technological innovation, and institutional innovation. From the viewpoint of scientific collaboration, a
complementary relationship exists between knowledge, expertise, and resources among the collaborative subjects. The scientific collaboration relationship is a symbiotic relationship because there is continuous exchange of information and transmission of knowledge, which evolves with the collaboration. This produces not only new knowledge units but also new symbiotic forms, thereby transforming the collaboration structure. Therefore, scientific collaboration is the practical response to the symbiotic relationship between the subjects.

Chunqing Yuan explicitly put forward the theory of symbiosis in 1998, which distinguishes four types of symbiotic relationships based on the subjects’ behavior patterns: parasitism, commensalism, symmetrical mutualism, and asymmetric mutualism. In this paper, we borrow the symbiosis degree index from symbiosis-degree analysis, which is generally applied in biology research to determine the behavior pattern between the selected symbiotic subjects (A and B) in different periods. To ensure a continuous and uninterrupted collaboration relationship between the two symbiotic subjects, the following two criteria were followed when choosing the subjects: (1) the two scientists collaborated for more than 15 years, and (2) they maintained close collaboration, that is, they published at least one co-author paper once a year. Based on this, we defined the scientific collaboration symbiosis degree \( R_i \) of symbiotic individuals in the period \( i \), which is represented as

\[
R_i = \frac{(y_{i+1} - y_i)/y_i}{(x_{i+1} - x_i)/x_i},
\]

where \( x_i \) is the number of joint achievements of the symbiotic subjects A during the collaboration period \( i \) (i.e., the number of collaborative papers published, the number of co-citations, etc.), and \( y_i \) is the total number of achievements of the symbiotic subjects B in the scientific research during the period \( i \) (total number of papers published, the number of citations, etc.). The formula reflects that the rate of change in all the achievements of a scientist (total number of papers published or total citations) depends on the rate of change of the joint contributions (the number of co-authored papers or joint citations) that are obtained through scientific collaboration. Therefore, it reveals the roles of scientific collaboration in terms of the academic output (number of papers) or influence (cited frequency) of an academician. The symbiosis degree contributing to the scientific collaboration indicator reflects the extent of the benefit of the collaboration to the scientific research symbiotic subjects as well as the extent of the dependence of the symbiotic subjects on the collaborative relationship. Here, the number of individual publications and the citation frequency represent an individual’s research productivity and influence respectively. When \( x_i \) represents the number of papers, \( R_i \) represents the degree to which the symbiotic relationship contributes to scientific productivity. When \( x_i \) represents the citation
frequency, $R_i$ represents the degree to which the symbiotic relationship contributes to scientific influence.

The extent of dependence of the two symbiotic subjects (A and B) on the scientific collaboration in different periods is calculated using the symbiosis degree index ($R_a$ and $R_b$, respectively). Table 2 presents the four types of symbiotic relationship established between A and B under different conditions.

<table>
<thead>
<tr>
<th>Numerical comparison</th>
<th>Symbiotic relationship</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_a &gt; 0$, $R_b &gt; 0$</td>
<td>Mutualism</td>
<td>When the symbiosis degree index is positive and the values are not equal, the two symbiotic subjects have an asymmetric mutualistic symbiotic relationship; if the values are equal, then they have a symmetrical mutualistic symbiotic relationship</td>
</tr>
<tr>
<td>$R_a = 0$, $R_b &gt; 0$ or $R_a &gt; 0$, $R_b = 0$</td>
<td>Commensalism</td>
<td>One of the two symbiotic subjects benefits from the collaborative symbiotic relationship, while the other does not</td>
</tr>
<tr>
<td>$R_a &lt; 0$, $R_b &gt; 0$ or $R_a &gt; 0$, $R_b &lt; 0$</td>
<td>Commensalism</td>
<td>One of the two symbiotic subjects benefits from the collaborative symbiotic relationship, while the other does not</td>
</tr>
<tr>
<td>$R_a &lt; 0$, $R_b &lt; 0$</td>
<td>No symbiotic relationship</td>
<td>The two symbiotic subjects do not depend on the collaboration, the collaboration is not beneficial to the symbiotic subjects, and a real collaborative symbiotic relationship almost does not exist</td>
</tr>
<tr>
<td>$R_a = 0$, $R_b &lt; 0$ or $R_a &lt; 0$, $R_b = 0$</td>
<td>No symbiotic relationship</td>
<td>Collaboration is not beneficial to the symbiotic subjects, and a real collaborative symbiotic relationship almost does not exist</td>
</tr>
<tr>
<td>$R_a = R_b = 0$</td>
<td>No symbiotic relationship</td>
<td>The two symbiotic subjects develop independently</td>
</tr>
</tbody>
</table>

In addition, owing to the limitations of the actual situation, it is unlikely that $R_a = R_b$. Under normal circumstances, there is probably a difference between the experimental value and the theoretical value; this error falls within a limited range, that is, we consider the experimental value and the theoretical value to be almost equal. In this paper, we borrow the relevant concepts of probability and statistics; moreover, when $0.95 < \frac{R_a}{R_b} < 1.05$, we consider that the values of $R_a$ and $R_b$ are relatively close and there is a symmetrical mutualistic symbiotic relationship between A and B.

**Empirical analysis**

*Evolution characteristics of the collaborative symbiotic network between a teacher and student*
Based on the collected data, we use the Thomson Data Analyzer (TDA) to clean the data, extract the author’s information regarding all the papers, and then create the collaboration matrix between the academicians after further processing. Ucinet is used for the network index analysis and for drawing the co-authorship network map at each time point. Figure 1 shows the collaboration network between academicians Xia Jianbai and Li Shushen. In this figure, we observe that the two academicians had their own collaboration networks in different periods and maintained a strong collaborative relationship. In Figure 1, the reason why such a division of the time period between teacher and student, based on before and after teacher and student elected as academicians, taking every five years for a stage. The teacher Xia Jianbai was elected as an academician in 2001, the student Li Shushen was elected as an academician in 2011, so we take 2001 and 2011 as the dividing point to explore the teacher-student symbiotic relationship changes before and after they elected as academicians.

Figure 1. Scientific collaboration network between Xia Jianbai and Li Shushen
After conducting centrality analysis of the collaborative relationship between the two academicians, the centrality of the network is obtained by ranking the top five co-authors, as shown in Table 3. After analyzing the data, we can conclude that in the collaboration network, the centrality degree of Li Shushen is higher than that of Xia Jianbai and that the closeness centrality and betweenness centrality of Xia Jianbai are higher than Li Shushen’s. Thus, it can be seen that young academicians have a certain advantage in the number of papers published; however, old academicians are stronger than the young academicians in the important role of the overall collaborative relationship and the utilization of the overall resources of these two aspects. At the same time, another researcher Li Jingbo maintained a close collaboration with these two academicians. Investigating Li Jingbo’s academic background, we found that he also worked at the China Institute of Semiconductors. In 2001, he was awarded his doctor’s degree from the China Institute of Semiconductors. In 2007, he was selected by the “Hundred Talents Program” of the Chinese Academy of Sciences. In 2009, he won the “National Outstanding Youth Fund” and many other honors.

Table 3. Centrality analysis of the two academicians’ scientific collaboration network

<table>
<thead>
<tr>
<th>The top five</th>
<th>Degree centrality</th>
<th>The top five</th>
<th>Closeness centrality</th>
<th>The top five</th>
<th>Betweenness centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li, Shushen</td>
<td>2.064</td>
<td>Xia, Jianbai</td>
<td>73.251</td>
<td>Li, Shushen</td>
<td>62.306</td>
</tr>
<tr>
<td>Xia, Jianbai</td>
<td>1.983</td>
<td>Li, Shushen</td>
<td>69.531</td>
<td>Li, Shushen</td>
<td>51.037</td>
</tr>
<tr>
<td>Li, Jingbo</td>
<td>0.657</td>
<td>Li, Jingbo</td>
<td>53.776</td>
<td>Li, Jingbo</td>
<td>1.643</td>
</tr>
<tr>
<td>Niu, Zhichuan</td>
<td>0.271</td>
<td>Niu, Zhichuan</td>
<td>52.353</td>
<td>Zhang, Ping</td>
<td>1.574</td>
</tr>
<tr>
<td>Wei, Suhuai</td>
<td>0.261</td>
<td>Wei, Suhuai</td>
<td>52.047</td>
<td>Chang, Kai</td>
<td>1.358</td>
</tr>
</tbody>
</table>

Collaborative symbiotic relationship between a teacher and student

Did the symbiotic relationship between the two academicians remain immutable in the past 20 years? To explore the evolutionary process of the symbiotic relationship between the two academicians, we compare the scientific collaboration symbiosis degree indices (number of co-authored papers and joint citation frequency). For Li Shushen and Xia Jianbai, we obtain the SCI datasets pertaining to their published papers and select all the SCI papers co-authored by the two academicians. Then, we examine the time-series distribution of the number of collaborative papers and the joint citation frequency of the two academicians for the last 20 years (Table 4). Both indices gradually increased, reaching a peak in 2006–2010; then, they decreased.

Table 4. Distribution of the number and the citation frequency of co-authored papers published by the two academicians

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of co-authored papers</td>
<td>2</td>
<td>8</td>
<td>16</td>
<td>47</td>
<td>17</td>
</tr>
<tr>
<td>Total number of papers of Li Shushen</td>
<td>2</td>
<td>12</td>
<td>48</td>
<td>80</td>
<td>65</td>
</tr>
<tr>
<td>Total number of papers of Xia Jianbai</td>
<td>7</td>
<td>47</td>
<td>40</td>
<td>78</td>
<td>25</td>
</tr>
<tr>
<td>Joint citations</td>
<td>30</td>
<td>322</td>
<td>505</td>
<td>1408</td>
<td>166</td>
</tr>
<tr>
<td>Total citations of Li Shushen</td>
<td>30</td>
<td>354</td>
<td>856</td>
<td>1759</td>
<td>174</td>
</tr>
<tr>
<td>------------------------------</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td>Total citations of Xia Jianbai</td>
<td>86</td>
<td>979</td>
<td>829</td>
<td>1669</td>
<td>172</td>
</tr>
</tbody>
</table>

Based on the above statistical results, the number of papers represents the scientific research productivity of the academician, and the citation frequency represents the scientific research influence. We calculate the symbiosis degree of scientific collaboration productivity and influence of the collaborative papers of the two academicians in five periods between 1994 and 2014 (Tables 5 and 6, respectively). In these two tables, $R_1$ represents the symbiosis degree of the scientific collaboration of Li Shushen, and $R_2$ represents the symbiosis degree of the scientific collaboration of Xia Jianbai.

Figure 2 shows the histogram of the symbiosis degree of collaborative productivity between teacher and student; the abscissa represents the time in years, and the ordinate represents the size of the symbiosis degree of productivity indicator based on the analysis of the symbiosis degree of collaborative productivity. In 1994–1995, $R_1 > R_2 > 0$, which reveals that there is an asymmetric mutualistic symbiosis relationship between the two academicians and that there is a large difference in terms of dependence on the collaboration between the two academicians. The student relied on the collaboration in the initial academic stage; collaborative papers contributed more to Li Shushen’s scientific productivity. In 1996–2000, $R_2 > R_1 > 0$, which indicates that the two subjects were both more dependent on the collaborative relationship; their relationship transformed to an asymmetric mutualistic symbiotic relationship. Although there are some differences between the two academicians in terms of the extent of their dependence on the collaboration, the differences is not too large, and their academic collaboration proceeded harmoniously. In 2001–2005, $R_1 > 0 > R_2$, which indicates that the collaborative relationship between the two academicians in this period was beneficial to the scientific research output of Li Shushen; however, there was no advantage for the scientific research output of Xia Jianbai. After the teacher was elected as an academician, the student became more dependent on collaboration with the teacher. Therefore, the symbiotic relationship between the two academicians at this time is of the form of commensalism. In 2006–2010, $R_2 > R_1 > 0$, which indicates that there was a certain dependence on the collaborative relationship between the two academicians at this time. There is a difference in the dependence on the collaboration between the two, but this is not large, and the degree of dependence is small; the relationship between the two is an asymmetric mutualistic symbiotic relationship. In 2011–2014, $R_2 > R_1 > 0$, which indicates that Xia Jianbai had a stronger dependence on the collaboration. There is a big difference between the two subjects in terms of the degree of dependence on the collaboration; the teacher had a strong dependence on the collaboration with the student in the later period of his academic career, and the relationship between the two is an asymmetric mutualistic symbiotic relationship.
Table 5. Symbiosis degree of collaborative productivity between teacher and student

<table>
<thead>
<tr>
<th>Period</th>
<th>Li Shushen (R_i)</th>
<th>Xia Jianbai (R_j)</th>
<th>Numerical comparison</th>
<th>Judgment of symbiotic relationship</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1994, 1995}</td>
<td>1</td>
<td>0.29</td>
<td>(R_1 &gt; R_2)</td>
<td>asymmetric mutualism</td>
<td>Li Shushen’s dependence on collaboration is greater than that of Xia Jianbai</td>
</tr>
<tr>
<td>{1996, 2000}</td>
<td>1.67</td>
<td>1.9</td>
<td>(R_1 &lt; R_2)</td>
<td>asymmetric mutualism</td>
<td>Xia Jianbai’s dependence on collaboration is slightly greater than that of Li Shushen</td>
</tr>
<tr>
<td>{2001, 2005}</td>
<td>3</td>
<td>-0.15</td>
<td>(R_1 &gt; R_2)</td>
<td>commensalism</td>
<td>Li Shushen’s dependence on collaboration is much greater than that of Xia Jianbai</td>
</tr>
<tr>
<td>{2006, 2010}</td>
<td>0.34</td>
<td>0.49</td>
<td>(R_1 &lt; R_2)</td>
<td>asymmetric mutualism</td>
<td>Xia Jianbai’s dependence on collaboration is slightly greater than that of Li Shushen</td>
</tr>
<tr>
<td>{2011, 2014}</td>
<td>0.29</td>
<td>1.06</td>
<td>(R_1 &lt; R_2)</td>
<td>asymmetric mutualism</td>
<td>Xia Jianbai’s dependence on collaboration is greater than that of Li Shushen</td>
</tr>
</tbody>
</table>

Figure 2. Symbiosis degree of collaborative productivity between teacher and student

Figure 3 shows the histogram of the symbiosis degree of collaborative influence between the teacher and student, where the abscissa indicates the time in years, and the ordinate indicates the value of the symbiosis degree of the influence indicator. In 1994–1995, \(R_1 > R_2 > 0\), which indicates that there is an asymmetric mutualistic symbiotic relationship between the two subjects. As in the case of the symbiosis degree of productivity of their collaborative papers, there is a big difference between the two subjects in terms of their dependence on the collaboration, and the citation frequency produced by collaborative papers contributes the most to the influence of the published papers of young Li Shushen. In 1996–2000, \(R_1 > R_2 > 0\), which indicates that the two subjects have greater dependence on the collaborative
relationship, i.e., $\left| \frac{R_2}{R_1} \right| > 0.95$. There is a difference between the two in terms of their dependence on the collaboration; however, this is only 0.003, close to the symmetric mutualistic symbiosis state. In 2001–2005, $R_1 > 0 > R_2$, which indicates that during this period, the collaborative relationship between the two was beneficial to the academic influence of Li Shushen, but there was no advantage to the academic influence of Xia Jianbai. After the latter was elected as an academician, his student became more dependent on their collaboration; their symbiotic relationship was of the form of commensalism. In 2006–2010, $R_1 > R_2 > 0$, which indicates that the two have a certain dependence on the collaborative relationship during this period, i.e., $\left| \frac{R_2}{R_1} \right| > 0.95$. The difference in the dependence between the two is 0.02, which is very small. Thus, we consider that their relationship is a symmetrical mutualistic symbiotic relationship. In 2011–2014, $R_1 > R_2 > 0$, which indicates that the subjects have a certain dependence on the collaborative relationship, i.e., $\left| \frac{R_2}{R_1} \right| > 0.95$. In terms of influence, there is a small difference of 0.04 between their dependence on the collaboration. Thus, their relationship may be characterized as a symmetrical mutualistic symbiotic relationship.

Xia Jianbai was elected as an academician in 2001. In 1996–2000, the two scholars maintained a high degree of symbiosis in their scientific collaboration, indicating that the two scholars who were elected as academicians were dependent on each other through scientific collaboration to improve their academic productivity and influence. In 2001–2005, Xia Jianbai was elected as an academician. Li Shushen had not been elected yet. At that time, Li Shushen relied more on Xia Jianbai, who was an academician, and their collaboration was more beneficial to Li Shushen, who was not elected as an academician.

<table>
<thead>
<tr>
<th>Period</th>
<th>Li Shushen $(R_1)$</th>
<th>Xia Jianbai $(R_2)$</th>
<th>Numerical comparison</th>
<th>Judgment of symbiotic relationship</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1994, 1995}</td>
<td>1</td>
<td>0.35</td>
<td>$R_1 &gt; R_2$</td>
<td>asymmetric mutualism</td>
<td>Li Shushen’s dependence on the collaboration is greater than that of Xia Jianbai</td>
</tr>
<tr>
<td></td>
<td>1.009</td>
<td>1.006</td>
<td>$R_1 &gt; R_2$, $\left</td>
<td>\frac{R_2}{R_1} \right</td>
<td>&gt; 0.95$</td>
</tr>
<tr>
<td>{2001, 2005}</td>
<td>1.6</td>
<td>−0.5</td>
<td>$R_1 &gt; R_2$</td>
<td>commensalism</td>
<td>Li Shushen’s dependence on the collaboration is much greater than that of Xia Jianbai</td>
</tr>
</tbody>
</table>
From the above-mentioned example of a collaborative symbiosis between a teacher and student, we can see that the change in the number of collaborative papers affected the scientific research productivity of the two academicians. The young academian was more likely to benefit in 2001–2005, whereas in the other periods, the relationship between them was an asymmetric mutualistic symbiotic relationship. When they were in the initial stage of their collaboration, the young academian was more likely to benefit. However, with time, in the later stage of their collaboration, the old academian relied on the young academian to some extent. The number and the citation frequency of their collaborative papers have different impacts on the total number and the total citation frequency of the research papers of the two academicians; the symbiotic relationship between the two academicians appears to have varied in degree over time.

**Conclusion**

In this paper, we examined scientific collaboration structure between a teacher and student from the viewpoint of symbiosis by constructing the scientific research collaborative symbiosis network between them and analyzing the evolution of their symbiotic relationship and its effect on their productivity and influence. The following conclusions are drawn from this empirical analysis: (1) In a collaborative
symbiosis network between a teacher and student, the old leads the young, i.e., the young researchers draw support from the teachers’ integrated resources to improve their levels of doing scientific research, stepping forward unceasingly and growing. (2) In a collaborative symbiosis network, apart from the two symbiotic subjects, the other collaborators play an important role through their scientific achievements and academic status. (3) The degree of collaborative symbiosis between the teacher and student has obvious characteristics of alienation with time; at the beginning of their collaboration, the student relies more on the collaboration, whereas later on, the teacher’s reliance on the collaboration becomes greater than the student’s.

This study can serve as a reference for exploring the structure and characteristics of the symbiosis network of scientific collaboration and optimizing the selection of scientific partners in the process of scientific research. Firstly, in the process of scientific research, we should pay attention to the collaboration between teachers and students; the teachers’ driving effect on the students should not be underestimated. Secondly, scientific celebrities can produce significant knowledge spillover effects, striving for more resources for the field they belong to and contributing to the overall output growth in this field, which facilitates the other scientists to engage in research more effectively. In future research, we believe that the symbiotic relationship between two collaborative symbiotic subjects can be extended to the study of the symbiotic relationship of the ternary structure, further expanding the research objects and scope of the study and providing more data on symbiotic relationships in scientific collaboration.

The limitations of this paper are as follows. Firstly, this paper only chose Xia Jianbai and Li Shushen as an example of the symbiotic relationship between teachers and students. In later studies, we will continue to expand the number of research sample and the research field to make the research results more general. Secondly, we will study the interdisciplinary or multidisciplinary collaboration between scholars contrasting with symbiotic collaboration research. Thirdly, this paper only explores the changes of the symbiotic relationship between teacher and student before and after they elected as academicians respectively. In future research, we will consider matching the symbiotic relationship types with career stages, for instance, the PhD period, the Postdoc period, the first years at Assistant Prof. level etc.

Acknowledgments
This research received the financial support from National Science Foundation of China: Study on the Structure and Evolution of the Scientific Collaboration Network of Academicians from the Perspective of Symbiosis: A Case Study of Academicians of CAS and NAS under grant number 71603015. Our gratitude also goes to the anonymous reviewers for their valuable comments.

References


A Brief Analysis of Top Scientists in the Field of Economics and Business Based on the Essential Science Indicators Database

Shanshan Wan¹  Nan Zhang²  Peiling Wang³  Peng Zhang⁴  Qiang Wu⁵

¹wanshan@mail.ustc.edu.cn
University of Science and Technology of China, Hefei (China)

²zntj2012@mail.ustc.edu.cn
University of Science and Technology of China, Hefei (China)

³peilingw@utk.edu
University of Tennessee at Knoxville, Knoxville (USA)

⁴zpbeidou@mail.ustc.edu.cn
University of Science and Technology of China, Hefei (China)

⁵qiangwu@ustc.edu.cn
University of Science and Technology of China, Hefei (China)

Abstract
Based on the Essential Science Indicators (ESI) database, this study analysed 2140 highly cited papers (HCP) and 4499 authors of these articles in the field of Economics and Business. By ESI definition, the HCP are those in the top 1% of the rankings. This study defines a 'top scientist' as one with at least 5 HCP and determined that there were 76 top scientists in the period 2005-2014, among which 46 researchers (60.5%) were from the USA. The results show a significant positive correlation between HCP and h-index of these top scientists. The study also found that the collaboration network among these top scholars is not very close, but a relatively tight subnetwork with 13 top scientists has formed.

Conference Topic
Studies on the level of individual scientists; The application of informetrics on evaluation

Introduction
Analysing the most influential scientists is one of the hot issues in research on the level of individual scientists. Many scholars have carried out extensive research into such topics including top scientists (Costas, van Leeuwen, & Bordons, 2012; Abramo, D’Angelo, & Soldatenkova, 2017; Charlton, & Andras, 2008; Baccini et al., 2012), core scientists (Furukawa & Goto, 2006), star scientists (Zucker & Darby, 1996), top researchers (Goodall, 2006; Miller, Coble, & Lusk, 2013; Lima et al., 2015; Kegen, 2015; Cortés, Mora-Valencia, & Perote, 2016), highly cited researchers (Clarivate Analytics, 2016), top authors (Walters & Arndt, 1999), and top scholars (Thieme, 2007). They have come to some valuable conclusions, but for the question of how to identify the top scientists, there are a variety of practices.

The current paper proposes that highly cited papers (HCP) can be used as a viable approach to distinguish the most influential scholars. In the Essential Science Indicators (ESI) database, HCP are those papers that rank in the top 1% by citation counts for field and year, using a time span limited to 10 years with bimonthly updates during the current year. The ESI database is based on over 10 million articles published in the more than 12000 journals included in the Web of Science database, and this database has become one of the most important tools to evaluate the academic influence of researchers, universities, institutions, and countries.
In this study, the ESI database was used to identify the top scientists in the field of Economics and Business. This paper proposes the criterion that top scientists should have at least five HCP. The paper is driven by the following questions:

Q1. According to the number of highly cited papers, who are the top scientists in Economics and Business? Which countries do they come from?

Q2. Is there a correlation between HCP and h-index of these top scientists?

Q3. Are there any collaboration networks among these top researchers?

**Methods and data**

We identified 2140 highly cited papers in the field of Economics and Business from 2005 to 2014 in the ESI database and then downloaded these papers from the Web of Science one by one. The ESI database is updated every two months. The data collection time was July 22 to 26, 2015, soon after the closest EIS update time on July 7, 2015. These 2140 HCP (two retracted publications were ignored) include 1899 articles (88.7%) and 241 reviews (11.3%). Figure 1 shows an increasing trend of HCP in Economics and Business during the period of 2005-2014. The number of highly cited papers increased from 140 papers in 2005 to 252 papers in 2014, an increase of 80%. The year that had the highest number of highly cited papers was 2013 with 259 papers, followed by 2012 with 257 papers, and 2014 with 252 papers. Regression analysis showed that HCP increased with the year of publication, with $R^2$ being 0.8863 and the regression model indicating significance ($p < .01$).

![Figure 1. Trends by year of highly cited papers in Economics and Business (2005-2014).](image)

When analysing the collaboration network among top scientists, this study used the software Sci². Based on Cyber Infrastructure Shell (CIShell), Sci² was developed by Katy Börner and his team in Indiana University (Sci² Team, 2009). As an open source software framework, CIShell is powerful for integrating datasets, algorithms, tools, and computing resources easily. Not only can Sci² determine statistics and carry out other analysis of data, but it can also detect many kinds of networks, for example, co-author networks and co-cited networks.

In calculating h-indices (Hirsch, 2005) of top scientists in Economics and Business, this paper used the Web of Science database. The scientists’ h-indices were collected in late October and early November of 2016. The data was from the Web of Science with the search tags of ‘AU = (a top scientist’s name) AND SU = (Business & Economics)’, and the time span was set as
2005 to 2014. After getting the scientists’ number of publications and citations and distinguishing the authors who have the same name, we could acquire their h-indices.

**Results**

**Ranking analysis**

Overall, there were 4499 authors who together published the 2140 highly cited papers. Considering that HCP means top 1% in citation rankings, this study looked for the top 1% of these HCP authors. It was determined that listed 76 scientists had at least 5 HCP (Table 1). Although the number of such scientists reached 1.6%, the lowest ranked author with 5 HCP is ranked 34th, which is in the top 1%. Thus, in this paper, a ‘top scientist’ is defined as one who had at least 5 HCP during the study period. J. J. Heckman, who won the Nobel Prize in 2000 for his pioneering work in econometrics and microeconomics, was the most productive scientist with 12 HCP, followed by A. Falk, L. Kilian, M. Lenzen, and M. W. Peng with 9 HCP each. There are 46 top researchers (60.5%) who come from the USA (i.e. doing their research in the USA). The Netherlands and England rank second with 5 top scientists each. Canada, Japan, and Hong Kong, China rank fourth with 3 top scientists each, followed by Australia, Austria, Germany, and Lithuania with two top scientists. Seven other countries, Belgium, Brazil, Finland, France, Italy, Norway, and Switzerland, have one top scientist each.

<table>
<thead>
<tr>
<th>Top scientist</th>
<th>Country</th>
<th>HCP No.</th>
<th>Rank</th>
<th>H-index</th>
<th>Top scientist</th>
<th>Country</th>
<th>HCP No.</th>
<th>Rank</th>
<th>H-index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heckman, JJ</td>
<td>USA</td>
<td>12</td>
<td>1</td>
<td>25</td>
<td>Chetty, R</td>
<td>USA</td>
<td>5</td>
<td>34</td>
<td>21</td>
</tr>
<tr>
<td>Falk, A</td>
<td>Germany</td>
<td>9</td>
<td>2</td>
<td>19</td>
<td>Diebold, FX</td>
<td>USA</td>
<td>5</td>
<td>34</td>
<td>15</td>
</tr>
<tr>
<td>Kilian, L</td>
<td>USA</td>
<td>9</td>
<td>2</td>
<td>21</td>
<td>Duflo, E</td>
<td>USA</td>
<td>5</td>
<td>34</td>
<td>19</td>
</tr>
<tr>
<td>Lenzen, M</td>
<td>Australia</td>
<td>9</td>
<td>2</td>
<td>20</td>
<td>Eisenhardt, KM</td>
<td>USA</td>
<td>5</td>
<td>34</td>
<td>14</td>
</tr>
<tr>
<td>Peng, MW</td>
<td>USA</td>
<td>9</td>
<td>2</td>
<td>27</td>
<td>Farh, L</td>
<td>Hong Kong, China</td>
<td>5</td>
<td>34</td>
<td>9</td>
</tr>
<tr>
<td>Gneezy, U</td>
<td>USA</td>
<td>8</td>
<td>6</td>
<td>20</td>
<td>Fujimori, S</td>
<td>Japan</td>
<td>5</td>
<td>34</td>
<td>6</td>
</tr>
<tr>
<td>Leuz, C</td>
<td>USA</td>
<td>8</td>
<td>6</td>
<td>19</td>
<td>Goto, M</td>
<td>Japan</td>
<td>5</td>
<td>34</td>
<td>22</td>
</tr>
<tr>
<td>Saez, E</td>
<td>USA</td>
<td>8</td>
<td>6</td>
<td>22</td>
<td>Graham, JR</td>
<td>USA</td>
<td>5</td>
<td>34</td>
<td>18</td>
</tr>
<tr>
<td>Shleifer, A</td>
<td>USA</td>
<td>8</td>
<td>6</td>
<td>26</td>
<td>Grant, AM</td>
<td>USA</td>
<td>5</td>
<td>34</td>
<td>28</td>
</tr>
<tr>
<td>Acemoglu, D</td>
<td>USA</td>
<td>7</td>
<td>10</td>
<td>29</td>
<td>Greenwood, R</td>
<td>Canada</td>
<td>5</td>
<td>35</td>
<td>15</td>
</tr>
<tr>
<td>Bloom, N</td>
<td>England</td>
<td>7</td>
<td>10</td>
<td>13</td>
<td>Gronroos, C</td>
<td>Finland</td>
<td>5</td>
<td>34</td>
<td>11</td>
</tr>
<tr>
<td>Chrisman, JJ</td>
<td>USA</td>
<td>7</td>
<td>10</td>
<td>23</td>
<td>Harrison, DA</td>
<td>USA</td>
<td>5</td>
<td>34</td>
<td>17</td>
</tr>
<tr>
<td>Gomez-Mejia, LR</td>
<td>USA</td>
<td>7</td>
<td>10</td>
<td>18</td>
<td>Hasegawa, T</td>
<td>Japan</td>
<td>5</td>
<td>34</td>
<td>5</td>
</tr>
<tr>
<td>Havlik, P</td>
<td>Austria</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>Helpman, E</td>
<td>USA</td>
<td>5</td>
<td>34</td>
<td>12</td>
</tr>
<tr>
<td>Hitt, MA</td>
<td>USA</td>
<td>7</td>
<td>10</td>
<td>33</td>
<td>Heyhoe, E</td>
<td>Australia</td>
<td>5</td>
<td>34</td>
<td>5</td>
</tr>
<tr>
<td>Pedersen, LH</td>
<td>USA</td>
<td>7</td>
<td>10</td>
<td>16</td>
<td>Jansen, JJP</td>
<td>Netherlands</td>
<td>5</td>
<td>34</td>
<td>11</td>
</tr>
<tr>
<td>Valin, H</td>
<td>Austria</td>
<td>7</td>
<td>10</td>
<td>7</td>
<td>Kyle, P</td>
<td>USA</td>
<td>5</td>
<td>34</td>
<td>10</td>
</tr>
<tr>
<td>Zavadskas, EK</td>
<td>Lithuania</td>
<td>7</td>
<td>10</td>
<td>26</td>
<td>Laeven, L</td>
<td>USA/Netherlands/England/Belgium</td>
<td>5</td>
<td>34</td>
<td>23</td>
</tr>
<tr>
<td>Aghion, P</td>
<td>USA</td>
<td>6</td>
<td>19</td>
<td>22</td>
<td>Lousnbery, M</td>
<td>Canada</td>
<td>5</td>
<td>34</td>
<td>14</td>
</tr>
<tr>
<td>Fischbacher, U</td>
<td>Switzerland</td>
<td>6</td>
<td>19</td>
<td>14</td>
<td>Lusch, RF</td>
<td>USA</td>
<td>5</td>
<td>34</td>
<td>19</td>
</tr>
<tr>
<td>Gabaix, X</td>
<td>USA</td>
<td>6</td>
<td>19</td>
<td>16</td>
<td>Luthans, F</td>
<td>USA</td>
<td>5</td>
<td>34</td>
<td>25</td>
</tr>
<tr>
<td>Hail, L</td>
<td>USA</td>
<td>6</td>
<td>19</td>
<td>8</td>
<td>Nelson, GC</td>
<td>USA</td>
<td>5</td>
<td>34</td>
<td>7</td>
</tr>
<tr>
<td>Hansen, PR</td>
<td>USA</td>
<td>6</td>
<td>19</td>
<td>12</td>
<td>Nunn, N</td>
<td>Canada/USA</td>
<td>5</td>
<td>34</td>
<td>13</td>
</tr>
</tbody>
</table>
As to the h-index (Table 1), the best performance was by M.A. Hitt (h-index = 33), followed by S.A. Zahra (31), D. Acemoglu (29), A.M. Grant (28), S.A. Zahra (27), M.W. Peng (27), A. Shleifer (26), E.K. Zavadskas (26), T. Sueyoshi (26), J.J. Heckman (25), H. Aguinis (25), and F. Luthans (25). In Figure 2, scatter plots between highly cited papers and h-index are shown, and a positive correlation between these two factors can be roughly judged. The Spearman correlation coefficient between them is 0.242 (p < .05), which indicates that when evaluating the influence of scholars, the HCP and h-index are mutually supportive indicators. They are considerably useful evaluation methods.

Collaboration network analysis

Figure 3 demonstrates that the collaboration network among the 76 scientists is quite disconnected. There are 43 subnetworks: one subnetwork with 13 nodes (i.e. top scientists),
one subnetwork with 7 nodes, two subnetworks each with 3 nodes, 11 subnetworks each with 2 nodes, and 28 isolated nodes.

Figure 3. The collaboration network of top 76 scientists in Economics and Business (2005-2014).

Clearly, the largest subnetwork is in the middle of the graph with 13 nodes, which includes P. Havlik (7 HCP), H. Valin, (7 HCP), A. Tabeau (6 HCP), M. Von Lampe (6 HCP), H. Van Meijl (6 HCP), D. Willenbockel (5 HCP), D. Van Der Mensbrugghe (5 HCP), P. Kyle (5 HCP), G.C. Nelson (5 HCP), T. Hasegawa (5 HCP), S. Fujimori (5 HCP), C. Schmitz (5 HCP), and E. Heyhoe (5 HCP). They are strongly interconnected and, among them, the most cooperation was between P. Havlik and H. Valin, who have published 7 HCP together. These 13 scholars come from many countries: Austria (P. Havlik, H. Valin), Netherlands (A. Tabeau, H. Van Meijl), France (M. Von Lampe), Japan (S. Fujimori, T. Hasegawa), USA (P. Kyle, G.C. Nelson), Germany (C. Schmitz), Italy (D. Van der Mensbrugghe), England (D. Willenbockel), and Australia (E. Heyhoe). Their main research topics are socioeconomic, agricultural economics, climate change, land-use change, ecosystems, and sustainability.

Conclusions

In this paper, based on ESI database, 2140 highly cited papers and 4499 authors in Economics and Business were analysed. In accordance with the HCP rankings in which ‘high’ means top 1%, this study identified the 76 ‘top’ scientists as those with at least 5 HCP. Of these top scholars, 46 (60.5%) were from the USA. The Netherlands and England rank second with five authors, respectively. The number of top scholars in the other 14 countries was below three. Moreover, the study shows that there is a fairly strong correlation between HCP and h-index. The Spearman correlation coefficient between them is 0.242 (p < .05), which indicates that when assessing the individual’s impact, the HCP and h-index are mutually supportive approaches.
This study also finds that the collaboration network among these top scholars is not very close, and there are 43 subnetworks. However, the analysis revealed a relatively tight subnetwork with 13 top scientists: P. Havlik (7 HCP), H. Valin, (7 HCP), A. Tabeau (6 HCP), M. Von Lampe (6 HCP), H. Van Meijl (6 HCP), D. Willenbockel (5 HCP), D. Van Der Mensbrugghe (5 HCP), P. Kyle (5 HCP), G.C. Nelson (5 HCP), T. Hasegawa (5 HCP), S. Fujimori (5 HCP), C. Schmitz (5 HCP), and E. Heyhoe (5 HCP). Most of their research focuses on socioeconomic, agricultural economics, climate change, land-use change, ecosystems, and sustainability.

Acknowledgments
This research was supported in part by the National Natural Science Foundation of China (Grants 71273250).

References


Measuring and analysing the internal, topical coherence of Web of Science Subject Categories

Stephan Stahlschmidt\textsuperscript{1} Marion Schmidt\textsuperscript{2}

\textsuperscript{1}stahlschmidt@dzhw.eu
German Centre for Higher Education Research and Science Studies (DZHW), Berlin (Germany)

\textsuperscript{2}schmidt@dzhw.eu
German Centre for Higher Education Research and Science Studies (DZHW), Berlin (Germany)

Abstract

The Web of Science subject categories constitute a ubiquitous standard partition of disciplines in the Web of Science database and many bibliometric analyses. Due to the classification of complete journal issues, they have been criticized as being too broad, while entailing substantial implications on normalization exercises. Several alternatives have been proposed, but due to the unknown ground truth no proposal has found widespread application. Furthermore an empirical analysis of the Subject Categories’ shortcomings on a comprehensive scale has not been conducted.

In this research-in-progress paper we thus measure the topical coherence of all Web of Science Subject Categories via an extensive clustering exercise including all articles indexed in the Web of Science between 2003 and 2012. This descriptive analysis will be supplemented by a regression analysis explaining the influence of several explanatory factors on the varying Subject Categories’ coherence levels.

Given the current state of the ongoing analysis, we can verify the coarseness of the Web of Science Subject Categories and additionally observe an increase in topical heterogeneity over time, which we will analyse in the upcoming regression.

Conference Topic

indicators, metrology, databases, methods and techniques

Introduction

Publication and citation indicators are usually normalized in order to compensate for different publication and citation behaviour. For bibliometric studies that are based on the Web of Science (WoS) the overlapping classification system of subject categories (SCs) as provided by the database producer is mostly used. Citation indicators such as the mean normalized citation score (Waltman, van Eck, van Leeuwen, Visser & van Raan, 2011) normalize the citation rate of a given publication body with expected values calculated on the basis of the subject categories. The observed values are therefore related to their subject contexts in order to ensure comparability.

The WoS SCs are often criticized within the bibliometric community, especially because the classification is not carried out at the level of the individual articles, but rather at the level of journals issues. This is generally considered to be relatively inaccurate, particularly regarding increasing interdisciplinarity (Gómez, Bordons, Fernández & Méndez, 1996). New journals are assigned to existing subject categories by partially manual, partially algorithmic methods on the basis of citation data (Leydesdorff & Rafols, 2009, Pudovkin & Garfield, 2002). These methods are not publicly documented.

In general, a wide range of alternative clustering and community detection algorithms is available (Kou, Peng & Wang, 2014, Subelj, van Eck & Waltman, 2016); theses algorithms can be based on a similarity measure between publications or entire journals. Different principles to infer the similarity have been proposed in the literature: they can be determined
either on the basis of citation data (usually bibliographic coupling or cross-citation) or the actual text (e.g. shared use of terms or phrases). In recent years, hybrid approaches combining citation and textual approaches were propagated (Thijs, Schiebel & Glänzel, 2013, Thijs, Zhang & Glänzel, 2015), as they compensate for both the lack of recall of the citation perspective and the lack of precision of the textual approaches.

While some comparative analyses of the WoS SC classification with alternative science maps based on inter-journal citation relationships have been proposed (Boyack, Klavans & Börner, 2005, Leydesdorff, 2006), these studies ultimately lack a gold standard to evaluate the diverse mappings. To partly address this issue and control for the randomness arising from the application of varying data sets, a comparative analysis of different methods on the basis of an identical data set is proposed in a current special edition of Scientometrics (Gläser, Glänzel & Scharnhorst, 2017). But without assured knowledge on the ground truth the benefit of entire alternatives to the proprietary SC classification (Glänzel & Schubert, 2003, Ruiz-Castillo & Waltman, 2014) cannot be assessed. Consequently none of these proposals have been adopted widely by the community until now, and the subject categories, probably due to their ubiquitous availability, are still the standard in the bibliometric community.

Nevertheless the subject categories of the WoS obviously vary in size and specificity. Van Eck, Waltman, Van Raan, Klautz, & Peul (2013) show evidence of heterogeneity within several medical subject categories along the dimensions of clinical and experimental research. Publications that are identifiable as experimental within the graphical representation exhibit higher citation counts. This finding poses, especially on the micro level, severe problems for the normalization with subject categories. However, due to its focus on medicine and a certain period of time, a comprehensive analysis on the subject is still missing, while being essential for an informed normalization in the computation of the MNCS and any percentile methods due to potential heterogeneous citation levels.

In this research-in-progress paper we analyse the topical coherence of the WoS SC classification via an internal clustering of articles within the SCs. By predominantly focusing on the subject-matter of an article we follow the general principle of scientific writing, namely progressing knowledge on a particular subject, and attach less value to diverging perspectives, which would result in a different partition of the articles (Wang & Rohe, 2016). The main research question is to what extent potential heterogeneities affect the normalization with SCs, whether more or less problematic areas of the classification with respect to coherence can be identified and how the phenomenon develops over time. Furthermore, we test approaches (such as citation autarky) to explain heterogeneous content structures with differing citation patterns within SCs in order to obtain informative background knowledge for any attempt to construct a more diverse alternative to the WoS SC classification.

**Method**

As stated, we assess the internal coherence of the WoS SCs by employing a hierarchical clustering approach and then by interpreting the number of resulting cluster as well as varying cluster specific citation levels as measures of coherence. Subsequently we analyse the observed level of coherence via explanatory factors. Consequently the clustering technique is not understood as resulting in a natural partition of the publications representing the unknown ground truth. It is rather applied as an operational measuring device (Dingle, 1950) to facilitate a valid comparison of the SCs’ internal coherence, because validating a representational measure with an unknown ground truth seems rather challenging (Hand, 1996). Indeed, the main interest of our work is a comparison of the operationally measured
coherence values across the SCs and a supplemental regression analysis explaining the influence of several explanatory factors on the varying SCs’ coherence levels.

To this end, we firstly define the relations among all articles in a particular SC and year according to their topical accordance and supply the resulting distance to the clustering algorithm. The subject-matter of an article is induced, albeit imperfectly, via the corresponding author keywords and reference list and we match any two articles based upon jointly applied author keywords and commonly referred literature (bibliometric coupling). Citations are predominantly understood as carrying cognitive information (Merton, 1957), marking concepts (Small, 1978) and consequently establishing a priority claim on the particular content (Kaplan, 1965). Given this perspective, citing a document might be understood as an act of acknowledgment and reward towards the cited document (Ravetz, 1971). Persuasive and perfunctory citations (Kaplan, 1965) diverge from this Mertonian ideal and illustrate the social facet of citations. In contrast, author keywords predominantly refer to concepts without any link to one or several persons. They thus reduce to some extent the social nuisance in this reference system.

A local cosine similarity $\text{sim}_{\text{loc}}^{(A,B)}_X$ between the articles $A$ and $B$ on the elements of the set $X \in \{\text{author keywords, referred literature}\}$ is computed separately for keywords and references via

$$\text{sim}_{\text{loc}}^{(A,B)}_X = \frac{|X(A) \cup X(B)|}{\sqrt{|X(A)||X(B)|}},$$

where $|\cdot|$ denotes the cardinality of a set. In a pre-processing step, the set of keywords is generalized by a stemming procedure to neutralize flexional variants used by the respective authors. At the same time the bibliometric coupling relation between any two articles is sharpened by excluding review articles, publications with a time lag of more than 10 years and non-source items, that is publications not published in the set of journals indexed in the WoS, but solely referred to by an indexed article.

Furthermore we extend this local perspective to a global level by transferring the local cosine similarity into a global one, $\text{sim}_{\text{glo}}^{(A,B)}_X$ (Ahlgren, Jarneving, & Rousseau, 2003). Thereby we consistently and throughout our study apply a global perspective including all publications in a particular SC and year at every step. We construct a similarity on such a global level by comparing the local similarities of article $A$ to all other articles $\hat{A}$ and itself with the vector of local similarities of article $B$ with $\hat{B}$ and itself. Apart from this theoretical argument of a consistent approach taking into account the whole SC, the global approach also incorporates important empirical advantages, like a decrease in the sparsity and a higher level of variance in the distance matrix decreasing the likelihood of ties and their severe consequences for clustering.

However, the cosine distance $\text{dist}_{\text{glo}}^{(A,B)} = 1 - \text{sim}_{\text{glo}}^{(A,B)}_X$ does not denote a proper distance metric, as neither the Cauchy–Schwarz inequality nor the coincidence axiom hold up. Hence, we transfer the cosine similarity to the angular distance via the inverse cosine and map the values in a normalization step into the interval $[0,1]$. Lastly we take an average on the distances resulting from author keywords and bibliometric coupling with equal weights in order to combine them to a hybrid distance measure (Braam, Moed & van Raan, 1991a, Braam, Moed & van Raan, 1991b)
These hybrid distances between all articles of a SC in a specific year constitute the basis for the subsequent clustering approach. We opt for an implementation based on agglomerative clustering foregoing the popular modularity based community detection algorithms due to their restriction to local and not global steps to improve the objective function. A local focus in the greedy optimization increases the likelihood to reach varying local optima across the SCs which would consequently diminish the comparability across SCs. In contrast, this comparability is ensured in our hierarchical clustering setting via a uniform application of the same linkage criteria and cut-off across SCs and years. We apply the so-called “ward linkage criterion”, following Bagatelj (1988), who justified the use of any dissimilarities instead of squared Euclidean distances for the Ward method.

Preliminary Results
Due to the extensive scope of the analysis incorporating all WoS SCs for the years 2003 - 2012, we currently can present preliminary results for a non-random sample of 19 SCs.

All SCs possess several clusters at some point of the observation period and can be classified as heterogeneous according to our measure. Taking into account the applied keywords and reference patterns, these SCs contain several subgroups differing by the respective subject-matter. The left panel of Figure 1 depicts the corresponding number of clusters per SC and year. The mean (90% quantile) number of clusters increases from 48 (122) to 95 (238) clusters in the observed time period, thus exhibiting a notable growth. This increase over time might mirror a potential rise in the level of specialisation in the WoS indexed articles, even when the actual amount of differentiation – measured by the number of clusters – depends on the applied configuration of both dissimilarity measures and clustering which might still be improved.

Whereas the observed increase in topical heterogeneity is interesting in itself, this scientific development entails a problem for bibliometrics only if the cluster specific citation levels vary within the observed clusters of a SC. The right panel in Figure 1 reports on the coefficient of variation on the cluster specific means:

\[ \text{CoV} = \frac{SD(\text{citations}_{\text{cluster}|\text{SC}})}{\text{citations}_{\text{SC}}}, \]

where \( \text{citations}_{\text{cluster}|\text{SC}} \) denotes a vector of average numbers of citations received by articles in the cluster \( c \in \{1, \ldots, C\} \) in a certain SC. The magnitude of the variation in mean citations across the clusters amounts on average to 75% to 150% of the mean citations received by the whole SC. This mean appears stationary without any trend, but is accompanied by a high volatility and demonstrates a substantial difference between the average citations of articles in the clusters and the SC as a total.

Give the current state of 19 SCs, the WoS SC classification seems too coarse to reflect the topical heterogeneity observed in the articles indexed in the WoS. Furthermore the heterogeneity seems to increase over time. However, a more detailed and comprehensive verdict will be reported once the processing of the remaining SCs has been completed.
Outlook

This research-in-progress paper will result in a descriptive analysis of the topical coherence of all SCs in the WoS for the years 2003-2012, an endeavour encompassing more than 10 million articles, and secondly will bring up explanatory models which explore the reasons for the observed variation in topical coherence across the SCs. Consequently we will present a descriptive analysis on the whole set of SCs and extract explanatory variables from the WoS database to conduct an inferential analysis which ideally results in informative background knowledge to construct a more diverse alternative to the WoS SCs classification.

References


Comprehensiveness and Overlap in Open Access Systems: Search Engines, Aggregate Institutional Repositories and Physic Open Sources

Ming-yueh Tsay¹ Tai-luan Wu²

¹mytsay@nccu.edu.tw
Graduate Institute of Library, Information and Archival Studies
National Cheng-Chi University (Taiwan)
Correspondence

²tailuanw@gmail.com
Graduate Institute of Library, Information and Archival Studies
National Cheng-Chi University (Taiwan)

Abstract
In this study, scholarly communication system of open access is examined through comprehensiveness and overlap of coverage. For open access systems, search engine includes Google Scholar; aggregate institutional repository include OAIster; the physic open source, i.e., arXiv. Noble laureates in physics from 2001 to 2013 are selected as samples in this study. Bibliographic records of their publications were retrieved and downloaded from each system, and a computer program was developed to perform the analytical tasks of sorting, comparison, elimination, aggregation and statistics. Bibliographic records retrieved from the three systems undertake quantitative analyses and cross references to determine the comprehensiveness and overlap of their system coverage. The results of the study may provide better references for scholars to select an appropriate open access system as an efficient scholarly communication channel; the academic institute may build institutional repositories, or independently create citation index systems in the future. Suggestions on searching enhancement for open access systems and our future work are presented accordingly.

Conference Topic
Open Access System, Comprehensiveness of System Coverage, Overlap of System Coverage, Scholarly Communication, Academic Assessment

Introduction

To assess the research performance, including research productivity and citation-based research impact, has become an important issue of scholars and research institutes. Take the prediction of the Nobel Prize winners as an example, Thomas Reuters, by using Web of Science, has studied and analyzed the citation results since 2002. They used it to identify the researchers who have the greatest influence in different disciplinary domains including chemistry, physics, psychology or medicine, economics, etc. to be the winners of the “Thomas Reuters Citation Laureates.” Since 2002, this prize has accurately predicted 27 Nobel laureates, indicating there is a certain overlap between the winners of the Citation Laureates selected by means of citation analysis and Nobel laureates. Hence, many people take the Citation Laureates as the leading indicator of Nobel Prize winners (Science Watch, 2013).

The rapid development of the Internet has facilitated the diversification of scholarly communication channels, not only altering the scholarly communication environment, but also further promoting the open access movement. At this stage, the traditional bibliometric study has evolved into webometrics. As the research data of webometrics come from the cyber world, websites, search engines and other open access systems, all can be research subjects or research tools of webometrics. How webometrics precisely and rigorously analyze and evaluate scholars, institutes and academic literature has been a major research topic; and as the important research foundation of webometrics, open access systems differ from one another with respect to the
objectives of construction, respective functions, and scope of collections. Therefore, in webometric studies, it is necessary to understand their differences if they are to be used as the research tools, which hence is the motivation of this research.

Under the “open access” environment institutional repositories, constructed by scholars and non-profit organizations such as academic institutes and libraries, may also possess the function of citation index. For example, the Cite Seer and Citebase in the computer science domain have constructed a citation index for its free online academic publications which include the arXiv.org that has archived the research literature in such domains as physics, mathematics, computer science, and biostatistics (Thelwall, 2008). Moreover, compared to commercial databases, search engines are one of the most popular tools for many users in searching useful information on the web. How should users choose between search engines and other open access systems? What are the differences among these sources? What are the pros and cons in terms of comprehensiveness and overlap of the collected data?

Based on the above background and motivations, this study compares the comprehensiveness and overlap of the data collected in open access systems using webometric methods. The open access systems selected in this study include search engines, Google Scholar; full-text institutional repositories, OAIster; and physics repositories, i.e. arXiv. The following objectives are expected to be achieved in this study:

The academic literatures published by the Nobel laureates in Physics are used as the research samples for comparing the comprehensiveness and overlap of the data collected in various open access systems.

Literature review

Bradford (1953, p. 154) first proposed research on literature overlap and uniqueness in 1937. His study focused on the overlap of journals collected by indexing and abstracting tools, i.e. cited literature simultaneously appearing in two or more databases. Based on the overlap theory by Bradford, Martyn (1967) analyzed the coverage of the abstract journals in his study and discovered that over half of the journals were collected in one or more index and abstract databases. He concluded this would result in a waste of manpower, as users would repetitively retrieve the same data as a result of the overlap in the index and abstract databases.

Nicholls (1989) compared four databases in the library and information domain, namely, Library and Information Science Abstracts (LISA), Educational Resources Information Center (ERIC), Library Literature, and Information Science Abstracts and retrieved 500,000 items of books and literary data during his research. He compared and explained the differences in the coverage among the four databases: Information Science Abstracts contained a large amount of seminar papers and special issue articles; LISA collected a large amount of foreign language materials and information science journals (with relatively few specialized journals on library and information;) the contents collected in Library Literature included books, analytical papers and a modicum of journals on library and information science, while the data collected in ERIC was the most complete among the four.

Hood (1998) conducted an overlap study of all the databases under the Dialog system using the theme of fuzzy set theory by downloading from all the databases of the Dialog system that included “fuzzy” data and selecting the literature related to the “fuzzy set theory.” Altogether he chose more than 30,000 data items between 1965 and 1993 from about 100 different databases. Hood removed the errors, unified all the fields and ultimately selected 15,644 items. He then conducted a comparative study of the overlap distribution, records of the most overlaps, sole records, the duplicate records within the databases, and the overlap among the top ten databases. The study found that there was great difference in the overlap records among the databases; of the 15,644 data items, 9,897 were only covered in one database,
accounting for 63.26%; the data collected by two databases numbered 1,922 pieces, accounting for 12.29%; and the data simultaneously collected in 12 databases numbered 5 pieces, taking up 0.03%. In addition, Hood found that there were internal duplicates in 28 databases, with MATHSCI having the greatest number of internal duplicates i.e. a total of 239 items. The reasons for the high overlap rate of internal duplicates were that both original articles and original papers with abstracts were covered in the database, and that the same article might be collected by various journals. The study finally explained the relative overlap among the various databases, pointing out the relative overlap in INSPEC and SCISEARCH exceeded 40%, which was the highest of all the databases in terms of overlap.

Zhu, Deng, Fang and Zheng (2008) conducted an overlap study on four search engines (i.e. Google, AltaVista, Alltheweb and WiseNut, a search engine in Korean) by using the top N (N=10, 20 or 80) retrieval results of 58 queries chosen from WordTracker service. The WordTracker service records the most popular queries submitted to some famous metasearch engines, e.g., MetaCrawler and Dogpile. The study indicated that the search results retrieved by the four search engines have little overlap. Google, on average, has the highest degree of overlap with the other search engines. Over 75% of the total distinct results are returned by only one search engine, and less than 3% are retrieved by all four search engines.

The study of Esmaeil, Kiaie and Ketab (2011) compared the overlap of six commonly used search engines systems and databases suggested by, Searchenginewatch.com. The study found that among the different search engine systems, the overlap rate of Yahoo with the other engines was the highest, which was about 40%. The recall rate of the physics journals of Curry Guide was 77.1%, and its overlap rate with the other five systems was 43.7%. Taking the physics domain as an example; the data in Meta Search Engine were the most complete among the six open access search engines.

Mitra and Awekar (2017) performed experiments on comparison overlap among four academic search engines (ASEs), Google Scholar, Semantic Scholar, Microsoft Academic and Scopus, which are computer science and engineering related. They collected approximately 2300 query terms from 2012 ACM Computing Classification System as well as 200 keywords from papers published in ACM SIGKDD 2016 Conference. Together, 2500 queries were sent to all four selected ASEs. The search results indicates that for each query, very few research articles appear in the top results list of all four ASEs. The disagreement of search results shows the overlap among of ASEs is significantly low.

In summary, overlap in the coverage and search results of traditional databases and web search engines is well studied. Most of these studies focusing on search engines, such as Google and Yahoo. There are some works investigating the indexing coverage of the web for major academic search engines. However, there is no study on the comparison of comprehensiveness and overlap among academic search engines, aggregate institutional repository and the subject related open access systems. Therefore, the present study is unique in its exploration of coverage of various open access systems on the basis of a quantitative bibliometric analysis of the literature in the subject area of physics.

Research methods

The present study compares the comprehensiveness and overlap of the data collected in various open access systems. Poyer (1984) defines the overlap of journal articles as the same journal article being collected and indexed or abstracted in two or more databases. Meanwhile, the rate of the same journal collected in two or more databases at the same time also constitutes overlap. On the basis of the suggestion of Poyer, in this study, overlap refers to the overlap rate of the same article being collected by two or more systems simultaneously.
This study collected the papers published by Nobel laureates in Physics (2001-2013) (see Appendix 1) and then produced a publication list to serve as the research sample for a comparison study of open access systems. The open access systems selected as research subjects were Google Scholar search engines, full-text institutional repositories, OAIster and physics disciplinary open access system arXiv.

Moreover, as the open access movement emerged in the 1990s, all open access systems, riding on this wave, were constructed in succession. The year of creation of the systems used in this study is respectively as follows: Google Scholar (2004), OAIster (2002) and arXiv.org (1991). Different construction years might have bearings on the difference in the quantity of the data collected in each system; the systems constructed earlier might cover data published in earlier years than those in the systems that were constructed later.

To create the research sample of this study – the publication list of the Nobel laureates in Physics, it was necessary to search the personal publications disclosed by these scholars before a publication list could be created. If the scholars did not provide their publication lists or the lists they provided were not complete, a further step of author search had to be taken to ensure the publication list of the research sample was complete and accurate.

This study searched the name of the aforementioned Nobel laureates in Physics in Google Scholar, OAIster, and arXiv.org separately. After authors with unique names were identified and eliminated, the retrieval results from the systems were then exported and collated. The following fields that had to be exported and analyzed in this study for identifying the correctness of the data in reviewing were the bibliographic records, title, author, publication year, publisher, journal title, volume and issue, number of pages etc.

To avoid errors in the bibliographic data occurred during massive download, the bibliographies were downloaded according to the complete descriptive format to create complete bibliographic data. The data processed by bibliometric methods mostly include text and numeric data. As different systems have different descriptive formats, incomplete descriptions and different description formats might occur when downloading bibliographic data. Therefore, the data must first be rectified through manual search to enable the original massive and jumbled bibliographic data to be standardized and the overlapping bibliographies retrieved from different databases to be removed through the Excel spreadsheet.

After the bibliographic data was collected and processed, manual verification and comparison were needed to ensure the accuracy of the research data, as there might be certain errors such as typographical errors in bibliographies, research papers unrelated to physics, alias of the author, and different titles and descriptive formats for the same journal. The bibliographic list created after eliminating the overlaps and undergoing the bibliographic data processing served as the publication list of the Nobel laureates collated with all the databases.

The completed publication list of the Nobel laureates in Physics must be cross-compared with the scholars’ publication information obtained from various Internet resources, the personal websites of the Nobel laureates in Physics, their research teams or the websites of the institutions they belong to. Only then could the research sample of the study be created – the publication list of the Nobel laureates in Physics (2001-2013).

Based on the research sample of the publication list of the Nobel laureates in Physics, the bibliographic titles were searched in the following three systems, Google Scholar, OAIster and arXiv.org. The search result of each title using each system was then individually recorded, including the descriptive errors and bibliography overlaps, which served as the basis for evaluating the comprehensiveness of and overlap in the bibliography of the Nobel laureates in Physics. Finally, the various results and review of the three systems were organized, statistically assessed, and compared which led to the conclusion and comments in this paper.
Research result

Through the research sample, the publication list of the Nobel laureates in Physics, this study hopes to understand the comprehensiveness and overlap of the literature on physics in the three open access systems under study. The following is a discussion and comparison of the comprehensiveness and overlap (including both internal and external overlaps) in the three open access systems.

Comprehensiveness of three open access systems

Based on the publication list of the Nobel laureates in physics, a total of 6094 items were searched in various open access systems, with the amount and comprehensiveness of bibliographic data shown in Table 1. Google Scholar had the most complete result with 5897 bibliographic items, covering 96.8% of the publication list; while OAIster which includes 45.3% of the data at the bottom. The physics open access system, arXiv includes 49% relevant items of the publication list. The extent of the arXiv collection could still not be compared to search engines, with a difference of about 48%. The comprehensiveness of institutional repository OAIster was not as good as arXiv.

As for the reasons for the highest comprehensiveness of data collected in Google Scholar, this might have been due to utilizing the various search engine to index the literature in the network world; its sources of literature include academic publishers, the institutional repositories of all universities and open access systems. With respect to this study, Google Scholar, in particular, has indexed the open access system of arXiv, rendering its bibliography the most complete.

For the foregoing, the comprehensiveness of the databases and systematically collected data evaluated in terms of physics literature was ranked as follows, starting with the highest: Google Scholar, arXiv, and OAIster.

<table>
<thead>
<tr>
<th>Open access system</th>
<th>Number of Bibliographic Items</th>
<th>Comprehensiveness Ratio*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Scholar</td>
<td>5,897</td>
<td>96.77%</td>
</tr>
<tr>
<td>arXiv.org</td>
<td>2,988</td>
<td>49.03%</td>
</tr>
<tr>
<td>OAIster</td>
<td>2,761</td>
<td>45.31%</td>
</tr>
</tbody>
</table>

* The total number of bibliographic items is 6,094.

Overlap of three open access systems

There are two types of system collection overlap: internal and external. The internal overlap, or self-overlap, occurred when duplicated bibliographies exist within each system. The external overlap, or relative overlap, is the same bibliographic items covered in two or more open access systems.

Internal overlap

The internal overlap of the bibliographic data from the retrieval of the list of works by the Nobel laureates in Physics from three open access systems is as shown in Table 2. The arXiv, which is physics disciplinary in nature, has lower internal overlap percentages than other open access systems. The arXiv has the least internal overlap i.e. no internal overlap (0%). Google Scholar has the highest internal overlap rate at 92.7%. As for the internal overlap rate of the institutional repository systems, OAIster is at 13.8%.
Table 2. Internal overlap of the publication list of the Nobel laureates in Physics in the open access systems

<table>
<thead>
<tr>
<th>Open access system</th>
<th>Bibliographic Items</th>
<th>Overlap Item</th>
<th>Overlap Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Scholar</td>
<td>5,897</td>
<td>5,464</td>
<td>92.66%</td>
</tr>
<tr>
<td>OAIster</td>
<td>2,761</td>
<td>381</td>
<td>13.80%</td>
</tr>
<tr>
<td>arXiv.org</td>
<td>2,988</td>
<td>0</td>
<td>0%</td>
</tr>
</tbody>
</table>

The reason for the high degree of internal overlap of Google Scholar is the Google Scholar’s indexing of the online world literature. For the publication list of the Nobel laureates in physics in Google Scholar, only 433 documents contained only one source while up to 92.7% of the bibliography contained two or more bibliographic sources. Moreover, Google Custom Search for repository content retrieval and searches the contents of all open access repositories. The retrieval results may include several bibliographic sources. Since the above-mentioned internal overlap is due to multiple bibliographic sources, the situation in each system may differ. Evidently, internal overlap records are caused by the title input discrepancies (e.g. the missing article in the title, the special symbols recorded in spelling, etc.) or the lack of bibliographic information (such as publication title, volume, pages, etc.). The difficulty of identifying one entry to another has deterred system to delete the duplicate bibliographic records.

**External overlap**

Table 3 demonstrates, compared with other open access systems in terms of overlap. Compared with other two open access systems, Google Scholar has overlap rates of 50.67%, 46.82%, respectively.

OAIster has an overlap rate of 100% with the Google Scholar search engine, showing that all the bibliographic data in OAIster are included in the Google Scholar. The arXiv shows the same situation.

The bibliographic overlap rate of OAIster in arXiv is 94.2%, while the overlap rate of arXiv in OAIster is 87.1%. This indicates the bibliographic overlap rate of these two open access systems is quite similar.

Table 3. External overlap of the publication list of the Nobel laureates in Physics in the open access systems (cross-column system and cross-reference system)

<table>
<thead>
<tr>
<th></th>
<th>Google Scholar</th>
<th>arXiv.org</th>
<th>OAIster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Scholar</td>
<td>X</td>
<td>50.67%</td>
<td>46.82%</td>
</tr>
<tr>
<td>arXiv.org</td>
<td>100%</td>
<td>X</td>
<td>87.08%</td>
</tr>
<tr>
<td>OAIster</td>
<td>100%</td>
<td>94.24%</td>
<td>X</td>
</tr>
</tbody>
</table>

To sum up, the overlap rate of arXiv, OAIster, compared with Google Scholar has reached 100%, suggesting the information found in other systems can be almost always obtained from Google Scholar. The reason for the high bibliographic overlap rate of Google Scholar with other open access systems is that Google Scholar has the most extensive bibliography and abundant information data resources. The overlap rates of arXiv and OAIster (87.1%) was slightly lower than that of OAIster and arXiv (94.2%). OAIster's bibliographic sources are mostly from academic institutions, which may result in bibliographic duplication overlap.

**Conclusion and suggestions**

Based on the results of the aforementioned comparison, this study analyzes the pros and cons of the three open access systems and makes recommendations according to the analysis
results to serve as a guide for scholars and researchers of webometrics for selection information systems.

In terms of the extensiveness and integrity of data regarding physics literature, the ranking is as follows: Google Scholar, arXiv and OAIster. Google Scholar has the most complete data coverage (96.8%), while OAIster’s integrity is the lowest (45.3%). For the physics open access systems, arXiv has the data coverage of 49%. The institutional repository system OAIster coverage is not as good as arXiv’s.

arXiv has the lowest internal overlap rate (0%) among the studied open access systems. The most internal overlap existing in Google Scholar is largely due to its various sources of bibliographic data, which made the bibliography under the system also differ (one journal from different bibliographic sources would show different index information). Since Google Scholar indexes the international literature online, 92.7% of entries contain two or more bibliographic sources.

The external overlap rate of arXiv, OAIster with Google Scholar was 100%. Over the course of the research, it was noted that Google Scholar’s main bibliographic data was from arXiv. This might have caused the 100% collection overlap of the data in arXiv’s, OAIster’s in Google Scholar.

For physics literature, a user is able to retrieve extensive data using a search engine such as the Google Scholar. In addition, Google Scholar provides users with links to the full-text of the journal (PDF files and web pages) if found. In addition to the scanned early physics literature for full-text search, arXiv also allows full access to open access journals and links to preprinted electronic documents. If these three open access systems do not provide the full text of the literature, the user should purchase commercial databases.

For the foregoing, the open access systems and the future work of this study need to be enhanced and improved as follows:

Open access systems need to implement bibliographic quality control and use a variety of screening mechanisms to eliminate unsuitable or incomplete bibliographic information to maintain the system data quality. The open access system also has to overcome the problems of invalid or incorrect links to avoid irrelevant search results, or the search result not being linked to the source site.

The future research on webometrics and collection overlap, if conducted with research methods such as data mining and/or text mining, will be beneficial to quantitative study. In this case, with query terms from controlled vocabulary system, e.g., subject thesaurus, or keywords from published papers are suggested as searching terms to retrieve relevant sets. Moreover, the future work of this study may be extended to include more academic open access systems and/or commercial databases, e.g. Wos or Scopus, for a complete comparison research.

Acknowledgments
This work was supported by grant NSC102-2410-H-004-221-MY2 from the National Science Council, Taiwan, R.O.C.

References


**Appendix**

**Appendix 1 Nobel Laureates in Physics (2001 – 2013)**

<table>
<thead>
<tr>
<th>Year</th>
<th>Names of Nobel Laureates</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>François Englert and Peter W. Higgs</td>
</tr>
<tr>
<td>2012</td>
<td>Serge Haroche and David J. Wineland</td>
</tr>
<tr>
<td>2011</td>
<td>Saul Perlmutter, Brian P. Schmidt and Adam G. Riess</td>
</tr>
<tr>
<td>2010</td>
<td>Andre K. Geim and Konstantin Novoselov</td>
</tr>
<tr>
<td>2009</td>
<td>Charles Kuen Kao, Willard S. Boyle and George E. Smith</td>
</tr>
<tr>
<td>2008</td>
<td>Yoichiro Nambu, Makoto Kobayashi and Toshihide Maskawa</td>
</tr>
<tr>
<td>2007</td>
<td>Albert Fert and Peter Grünberg</td>
</tr>
<tr>
<td>2006</td>
<td>John C. Mather and George F. Smoot</td>
</tr>
<tr>
<td>2005</td>
<td>Roy J. Glauber, John L. Hall and Theodor W. Hänsch</td>
</tr>
<tr>
<td>2004</td>
<td>David J. Gross, H. David Politzer and Frank Wilczek</td>
</tr>
<tr>
<td>2003</td>
<td>Alexei A. Abrikosov, Vitaly L. Ginzburg and Anthony J. Leggett</td>
</tr>
<tr>
<td>2002</td>
<td>Raymond Davis Jr., Masatoshi Koshina and Riccardo Giacconi</td>
</tr>
<tr>
<td>2001</td>
<td>Eric A. Cornell, Wolfgang Ketterle and Carl E. Wieman</td>
</tr>
</tbody>
</table>

MFTACA: An Author Co-citation Analysis Method Combined with Metadata in Full Text

Yi Bu\textsuperscript{1} Binglu Wang\textsuperscript{2} Win-bin Huang\textsuperscript{3,\ast} Shangkun Che\textsuperscript{4}

\textsuperscript{1} buyi@iu.edu
Indiana University Bloomington (America)

\textsuperscript{2} maplewang@pku.edu.cn
Peking University (China)

\textsuperscript{3} huangwb@pku.edu.cn
Peking University (China, corresponding author)

\textsuperscript{4} csk@pku.edu.cn
Peking University (China)

Abstract
As a frequently used method of depicting scientific intellectual structures, author co-citation analysis (ACA) has been applied to many domains. However, only count-based information is involved as the input of ACA, which is not sufficiently informative for knowledge representations. This article catches several metadata in full text of citing papers but not aims at content-level information, which could increase the amount of information input to ACA but not increase computational complexity a lot. By involving information including the number of mentioned times in a citing paper, the number of context words in a citing sentence, and the published year of a reference, we propose a new method called MFTACA (metadata-in-full-text-based ACA). We combine these pieces of information into traditional ACA and compare the results between ACA and MFTACA by using factor analysis, network analysis, and MDS-measurement. The result of our empirical study indicates that compared with traditional ACA, our proposed method shows a better clustering performance in visualizations and reveals more details in displaying intellectual structures.

Conference Topic
Citation and co-citation analysis; Mapping and visualization; Science of science

Introduction
Author co-citation analysis (ACA) is a bibliometric method in knowledge representation and has shown a good performance in depicting scientific intellectual structures and mapping knowledge domains (White and Griffith, 1981; McCain, 1990; Jeong, Song, and Ding, 2014). More than three decades since its born, ACA has been applied to many disciplines, such as library and information science (White & Griffith, 1998; Ding, Chowdhury, & Foo, 2001; Ding, 2011a; Zhao & Strotmann, 2014), cognitive science (Bruer, 2005), management science (Eom, 1999; Chen & Lien, 2011), and medical science (Chu, Liu, & Tsai, 2012).

Traditional ACA regards two authors with higher co-citation frequency as higher topical relatedness. Such assumption hints that every author pair with the same co-citation frequency as identical, which simply considers count-based instead of content-based information. As the availability of full-text data nowadays, Jeong \textit{et al}. (2014) firstly proposed content-based ACA method and compared the similarity between citing sentences. However, the computational cost of their content-based method could be high because of the processing of words as well as similarity calculation between citing sentences. Actually, with the full-text data, we do not have to employ the content-level information. Nevertheless, instead, several pieces of useful information at metadata level that were ignored previously in full text could be considered to improve the performance of ACA in mapping knowledge domains.
The number of mentioned times of references is a typical piece of information. As pointed out by Ding, Liu, Guo, and Cronin (2013) as well as Zhao, Cappello, and Johnston (2017), the number of mentioned times of a reference represents the importance of the reference to the citing paper. However, traditional ACA regards two co-cited authors with a distinct number of mentioned times in a citing paper as identical, which could be problematic. For example, Zhao and Strotmann (2014) cited (a) Zhao and Strotmann (2008a), (b) White and McCain (1998), and (c) Hirsch (2005), but (a) was mentioned 15 times, (b) twice, while (c) only once in the citing paper. The co-citation strength of pair (a)-(b) and (b)-(c) should be different when we consider their topical relatedness. In this paper, we start to consider the number of mentioned times of references as a supplement into ACA in order to provide more accurate information for mapping knowledge domains.

Besides, the citing sentences containing references could have different numbers of words. From our intuitive thinking, a reference contained in a longer citing sentence should have more topical relatedness to the citing paper than that is contained in a shorter one, because longer sentences are more likely to include more details or interpretations to the reference, which is useful to the citing paper—otherwise it is not likely to be cited with many interpretative words. However, traditional ACA ignores such difference in the length of citing sentences; as a result, we start to consider the difference in the length of citing sentence and combine it into ACA in this paper.

Moreover, the published time of references could also reveal the difference between co-cited authors. For instance, as pointed out by Bu, Liu, and Huang (2016), small difference between the published time of two references implies that the authors tend to focus on similar issues in the same period of time; therefore, the two authors’ relationship could be distinct in knowledge domain maps because they tend to use “various concepts, methods, or even diversified demands” (p.144) in dissimilar periods of time. The information of references; published time is considered in this paper and it is combined with traditional ACA to enhance the performance of knowledge domain mappings.

This article is outlined as follows. At first, the work related to our study and the data with the methods for our analysis are detailed. The findings as well as the comparisons between traditional ACA and our proposed MFTACA then are presented. Finally, the conclusions and the future research are pointed out.

**Related Studies**

Author co-citation analysis (ACA) was proposed by White and Griffith (1981). In 1990, McCain gave a completed overview and set up a standard framework for ACA, in which four steps of ACA implementations were mentioned: (1) Data collection and processing; (2) Construction of raw co-citation matrix; (3) Transformation to correlation matrix; and (4) Data analyses (e.g., factor analysis, clustering analysis, multi-dimensional scaling (MDS) analysis, and network analysis) and result interpretations. More than thirty years, this method has been improved a lot by revising rules to construct raw co-citation matrix and transform to correlation matrix. As pointed out by Persson (2001), the elements in co-citation matrix could be defined as first-author co-citation frequency or all-author co-citation frequency; the latter could to some extent provide more detailed knowledge domain maps (Rousseau & Zuccala, 2004; Zhao, 2006). Meanwhile, the rules of defining main diagonal values (Mêgnighbê to, 2013) and transforming correlation matrix (Ahlgren, Jarneving, & Rousseau, 2003; White, 2004) were both discussed and improved in detail to make the data processing more accurate. Additionally, several
metadata of references, such as published time and venue of references and their keywords, were considered in ACA implementations (Bu et al., 2016), and they were found to play positive roles in improving the performance of ACA maps. Note that the metadata they employed had been obtained from reference lists instead of full text.

Due to the availability of full-text scientific data, Jeong et al. (2014) first explored content-based ACA by comparing the similarity between citing sentences. Their empirical studies show that content-based ACA is able to mine more details in scientific intellectual depicting compared with traditional ACA. However, the computational complexity is high in full text processing, which impedes the applications of their proposed method to various domains widely.

However, when full-text data are used, it is not required to make content- or semantic-level analyses. Several non-content-level metadata are easily accessible in full text, such as the number of mentioned times and the number of context words in the citing paper. In addition, these pieces of information reflect the importance of reference to citing paper. For example, if mentioned many times than others, a reference could have more topical relatedness to the citing paper; if the citing sentence containing certain reference is longer than that containing other references, it is more likely to be interpreted more detailedly and thus has higher possibilities to be related to the citing paper. As a result, this paper combines the number of mentioned times and the number of context words of references into ACA and proposes a new method called MFTACA so as to improve the performance of ACA in knowledge domain mappings.

**Methodology**

The whole process of our algorithm is shown in Figure 1. All of the dataset are derived from full-text in Journal of the American Society for Information Science and Technology (JASIST, currently named as Journal of the Association for Information Science and Technology). After data processing (see details in “Data” section), we extract the number of mentioned times, the number of context words, and published years of references, and combine them into ACA (dotted area in Figure 1, see details in “Methods” section). After the new co-citation matrix is constructed, Pearson’s $r$ is utilized to transform it to correlation matrix. Factor analysis, network analysis, and MDS-measurement is used to process the data, in which Gephi (Bastian, Heymann, & Jacomy, 2009) is applied to display the results of the combined author co-citation network for discussion and analysis. Note that the dotted area in Figure 1 is the major difference among the proposed MFTACA and traditional ACA methods.

**Data**

The dataset used in this research is the same as that in Jeong et al. (2014), in which 1,420 full-text articles with citation links published in JASIST between January 2003 and June 2012 are selected. These 1,420 articles containing 60,068 references are completed by 32,095 authors. In order to make the co-citation matrix denser, we extract most popular 500 authors who have received the most number of citations since they are the most “popular” scholars and are often regarded as the representative of the research going back (Zhao & Strotmann, 2008a; Zhao & Strotmann, 2014).
Methods

Calculation of mentioned time parameters
A paper citing other papers refers that these cited papers (references) are related and useful to the citing paper (CP). Traditionally, these co-cited papers are regarded as equal in traditional co-citation analysis (Ding et al., 2013). However, the importance of cited papers could be distinct (Cano, 1989; Case & Higgins, 2000). Specifically, some references are crucial to the CP because they might be the foundation of CP. In the full-text, the important references could be revealed as multi-mentioned cited papers. For one CP, from an intuitive thinking, the more number of mentioned times a reference has, the more importance it is to the CP (Ding et al., 2013; Zhao et al., 2017). For instance, Zhao and Strotmann (2014) cited (a) Zhao and Strotmann (2008a), (b) White and McCain (1998), and (c) Hirsch (2005), but (a) was mentioned 15 times, (b) twice, while (c) only once. In this case, (a) should be the most important reference to the CP, compared with (b) and (c). Indeed, Zhao and Strotmann (2014) tried to map the knowledge domains of information science (IS) between 2006-2010 while (a) did the same thing but between 1996-2005 with similar methods, ACA and author bibliographic coupling analysis (ABCA), which shows that (a) is closely-related and important to the CP. However, (b) only used ACA to map the knowledge domain of IS instead of ABCA and the result of (b) is also very different from (a) and the CP. The reason (c) is cited is nothing but simply because the indicator “h-index” was used. Hence, we can see that the number of mentioned times of a reference is indeed positively related to its relatedness with CP.

Similar in ACA, if two co-cited authors are both mentioned many times in a CP, their topic relatedness could be high because both of them are closely related to the CP. In the above example, the topical relatedness of (a)-(b) should be higher than that of (b)-(c). Indeed, (a) and
(b) both focus on mapping IS with ACA, but (c) simply introduces an indicator to evaluate scholars. Thus, in our proposed algorithm, we assume that two co-cited authors with more number of mentioned times should be assigned more co-citation weights because they have higher possibilities to be related with each other topically.

Mathematically, suppose that author $A_i$ and $A_j$ are co-cited for $x_{ij}$ times. Specifically, the CP is annotated as $P_{ij,1}, P_{ij,2}, \ldots, P_{ij,x_{ij}}$. In $P_{ij,k}$ ($k = 1, 2, \ldots, x_{ij}$), assume that author $A_i$ is mentioned for $\lambda_{ik}$ times and author $A_j$ is mentioned for $\lambda_{jk}$ times. If we annotate the mentioned time of the cited author with the maximum number of mentions in paper $P_{ij,k}$ as $\lambda_{k_{\text{max}}}$, the mentioned time parameter between author $A_i$ and $A_j$ in paper $P_{ij,k}$, $MT_{ij,k}$, is calculated as:

$$MT_{ij,k} = \frac{\lambda_{ik}\lambda_{jk}}{\lambda_{k_{\text{max}}}}$$

(Eq. 1)

If we consider all of the $MT_{ij,k}$ in citing papers $P_{ij,1}, P_{ij,2}, \ldots, P_{ij,x_{ij}}$, the mentioned time parameter between author $A_i$ and $A_j$ among dataset, $MT_{ij}$, could be defined as:

$$MT_{ij} = \sum_{k=1}^{x_{ij}} MT_{ij,k}$$

(Eq. 2)

**Calculation of context word parameters**

When citing references, CPs tend to use one or more sentences to set up an argument, which is called citing sentences (or citance (Jeong et al., 2014)). However, the length of citing sentences could be distinct. For example, Zhao and Strotmann (2014) cited: (d) Finlay, Sugimoto, Li, and Russell (2012), (e) Milojević, Sugimoto, Yan, and Ding (2011), as well as (f) Sugimoto, Li, Russell, Finlay, and Ding (2011) in the same sentence with 27 words. They shared these 27 words and each of them has been assigned nine words averagely. Meanwhile, Zhao and Strotmann (2014) also cited (g) Zhao and Strotmann (2008b) in a sentence with 31 words. Although all of these references are cited once in the CP, their numbers of context words assigned could be diverse, 9, 9, 9, and 31, respectively.

Basically the number of context words assigned in a CP could reflect the importance of the reference. Specifically, more numbers of context words assigned in a CP reveal that more details and interpretations of the reference is likely to be stated, which hints that it has higher topical relatedness to the CP. For example, CP uses ACA and ABCA to map the knowledge domain of IS field, while (g) explores AACA at a methodology level; both of them are closely related to ACA research. Nevertheless, (d) analysed Library Science (LS) using titles and keywords, (e) employed article title words to depict the structure of LIS, and (f) focused on North American LIS dissertation using Latent Dirichlet Allocation (LDA) method, all of which are not as close as (g) in terms of the topical relatedness with the CP from an intuitive perspective. Indeed, the number of context words assigned to (g) is much more than that to (d), (e), and (f). These show that the number of context words assigned in a CP could be positively related to its topical relatedness to the CP. Similarly in ACA, if two co-cited authors are both assigned many words than others, they topical relatedness should be higher because both of them are closely related to CP.

In $P_{ij,k}$, assume that during its $\mu$th mention ($\mu = 1, 2, \ldots, \lambda_{ik}$), the citing sentence containing author $A_i$ ($A_j$) include $w_{ik\mu}$ ($w_{jk\mu}$) words and mentions $a_{ik\mu}$ ($a_{jk\mu}$) distinct authors ($w_{ik\mu}, w_{jk\mu}, a_{ik\mu}, a_{jk\mu} > 0$). The number of context words of author $A_i$ in paper $P_{ij,k}$, $cw_{i,k}$, could be calculated as (similar to $A_j$):

$$cw_{i,k} = \sum_{\mu=1}^{\lambda_{ik}} \frac{w_{ik\mu}}{a_{ik\mu}}$$

(Eq. 3)
If we annotate the largest number of context words of cited author in paper $P_{ij,k}$ as $cw_{k,max}$, the context word parameter between authors $A_i$ and $A_j$ in paper $P_{ij,k}$, $CW_{ij,k}$, is calculated as:

$$CW_{ij,k} = \frac{cw_{i,k}cw_{j,k}}{cw_{k,max}^2}$$  \hspace{1cm} (Eq. 4)

If we consider all of the $CW_{ij,k}$ in citing papers $P_{ij,1}, P_{ij,2}, \ldots, P_{ij,x_{ij}}$, the context word parameter between authors $A_i$ and $A_j$ among dataset, $CW_{ij}$, could be defined as:

$$CW_{ij} = \frac{1}{x_{ij}} \sum_{k=1}^{x_{ij}} CW_{ij,k}$$ \hspace{1cm} (Eq. 5)

### Calculation of published year parameters

Besides the mentioned time and context word parameters, we also employ published year parameter proposed by Bu et al. (2016), in which it is called time-based parameter. According to Bu et al. (2016), a small difference between two references’ published time refers that they tend to focus on similar research topics in the same time period; a large difference between two references’ published time refers that the representation of the co-cited authors’ relationships should be distinct because they are likely to use different methods, concepts, and tools although they might have similar research issues.

Suppose that $P_{ij,k}$ cites $r_{ik}$ papers published by author $A_i$ and $r_{jk}$ papers published by author $A_j$. The published year of these $r_{ik}$ papers published by author $A_i$ and $r_{jk}$ papers published by author $A_j$ is respectively $t_{ik1}, t_{ik2}, \ldots, t_{ikr_{ik}}, t_{jk1}, t_{jk2}, \ldots, t_{jkr_{jk}}$. Then the average of published year of author $A_i$ in the article $P_{ij,k}$ is calculates as:

$$TB_{i,k} = \frac{1}{r_{ik}} \sum_{v=1}^{r_{ik}} t_{ikv}$$  \hspace{1cm} (Eq. 6)

The published year parameter between authors $A_i$ and $A_j$ in paper $P_{ij,k}$ is defined as the sum of the average of published year of authors $A_i$ and $A_j$:

$$TB_{ij,k} = \frac{1}{1+ln(1+|TB_{i,k}-TB_{j,k}|)}$$ \hspace{1cm} (Eq. 7)

As a result, the average of published year of cited papers written by author $A_i$ among dataset is calculated as:

$$TB_{ij} = \sum_{k=1}^{x_{ij}} TB_{ij,k}$$  \hspace{1cm} (Eq. 8)

### Construction of the co-citation matrix based on three above parameters

The co-citation matrix in our proposed algorithm is based on three above parameters. All of the parameters are normalized into $[0,1]$. If we annotate the largest co-citation frequency among the dataset regardless of which author pairs as $x_{max}$, we can construct the new co-citation matrix, $M = (m_{i,j})$, as:

$$m_{i,j} = w_c \cdot \frac{x_{ij}}{x_{max}} + w_{MT} \cdot MT_{ij} + w_{CW} \cdot CW_{ij} + w_{TB} \cdot TB_{ij}$$  \hspace{1cm} (Eq. 9)

where the four positive weights, $w_c, w_{MT}, w_{CW}$, and $w_{TB}$, are relatively the weight of co-citation, mentioned time, context word, and published year parameters. Note that $w_c + w_{MT} + w_{CW} + w_{TB} = 1.0$.

### Results and Discussion

#### Factor Analysis

In factor analysis, we extract the factors whose Eigen factor is 1.0 or more as the result of factor analysis, regardless in ACA or MFTACA. As shown in Table 1, the number of factors extracted from MFTACA is 18, which is seven more than that in ACA. In terms of the total variance
explained, the factor analysis of ACA could explain 82.8% of total variance while that of MFTACA explains 86.1%.

Table 1. Factor analysis result overview.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Number of factors extracted</th>
<th>Total variance explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACA</td>
<td>11</td>
<td>0.828</td>
</tr>
<tr>
<td>MFTACA</td>
<td>18</td>
<td>0.861</td>
</tr>
</tbody>
</table>

Note: Here, \( w_c = 0.7, w_{MT} = 0.1, w_{CW} = 0.1, w_{TB} = 0.1 \), which are finally determined after examining lots of possible experiments. The same below.

Based on some previous research (Janssens, Leta, Glänzel, & Moor, 2006; Yang, Han, Wolfram, & Zhao, 2016), several core sub-fields of information science are dug out and more details are supposed to be refined. Table 2 shows the factor analysis results of both ACA and MFTACA. The same below.

Table 2. Factor analysis results.

<table>
<thead>
<tr>
<th>ID</th>
<th>Factor</th>
<th>ACA</th>
<th>MFTACA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Information retrieval</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>Information seeking behavior</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>3</td>
<td>Information usage, digital library</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>Language model, query, clustering</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>5</td>
<td>Classification algorithms, information organizations</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>6</td>
<td>Text mining, machine learning</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>7</td>
<td>User interface</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>8</td>
<td>User acceptance of information technology</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>9</td>
<td>Information systems</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>10</td>
<td>Data Mining, Data Analysis</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>11</td>
<td>Evaluation indicator, index</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>12</td>
<td>Webometrics, social network analysis</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>13</td>
<td>Visualization, mapping</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>14</td>
<td>Modeling</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>15</td>
<td>Scholarly communication</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>16</td>
<td>Journal citation analysis, interdisciplinarity, evaluation of algorithms</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>17</td>
<td>Network analysis</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>18</td>
<td>Bioinformatics</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Although using the same dataset as Jeong et al. (2014), we input approximately 500 authors into factor analysis while Jeong et al. (2014) did 100. We try to compare our results with theirs, which shows in bold in Table 2, where we can find that the results are similar and confirm each other’s. On the other hand, with respect to the factor analysis result of MFTACA, we can find...
many detailed sub-fields of information science, such as visualization and data mining. These newly-detected domains could showcase the nuance and emerging topics of information science recently. Therefore, we believe that MFTACA could reveal more details and nuance in depicting scientific intellectual structures.

**Network Analysis**

Figures 2 and 3 show the scientific intellectual structures by using the two methods, ACA and MFTACA, where each node represents an author and the size of the node is proportional to the degree of the node in the given network. The distance between nodes are determined by ForceAtlas2 (Jacomy, Venturini, Heymann, & Bastian, 2014), a frequently used layout algorithm in Gephi. If two nodes lie near in the map, for instance, their relationship could be strong; and *vice versa*. For visualization, we employ Modularity algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008), in which the nodes (authors) within the same colour indicate that their research interests are similar, while those in different colors show that their research interests could be distinct. The labels of the clusters are manually given by our reading literatures of the authors as well as browsing their personal websites. From Figure 2 we can see that four clusters are detected, bibliometrics, information retrieval, information behaviour, and library science/qualitative research. The results are similar to Jeong *et al.* (2014)’s result, where they also found four clusters, bibliometrics, information retrieval (I), information retrieval (II), and library science. On the other hand, Figure 3 detects six clusters; besides four clusters having been detected in Figure 2, it also finds two additional clusters, i.e. text mining/data mining and network-based information science. These two clusters reveal the nuance of the development of IS and are detailed sub-fields in IS. These indicate that our proposed MFTACA method could provide more details in knowledge domain maps and help better understand the domain.

![Figure 2. Knowledge domain map of traditional ACA method.](image)
Intuitively, the nodes within the same cluster lie nearer, and the nodes in different clusters lie farther in Figure 3 than in Figure 2. These indicate a better clustering performance in MFTACA method. Take Drs. K. W. McCain and R. Rousseau as examples. Both of them focus on bibliometrics during their main scientific careers. Specifically, they are both interested in ACA, where Dr. McCain (1990) gave a comprehensive overview of ACA and used ACA to map knowledge domain of IS field between 1972 and 1995 (White & McCain, 1998). Dr. Rousseau, as a representative in typical European bibliometricians, proposed several types of ACA by classifying them according to distinct requirements (Rousseau & Zuccala, 2004) and discussed whether Pearson’s $r$ should be used in ACA (Ahlgren, Jarneving, & Rousseau, 2004). Their positions in Figure 3 are nearer than in Figure 2, which shows that MFTACA could play a role of closing authors with similar research interests. Another examples come from Drs. A. Spink and B. Shneiderman, in which the former researcher is an expert in information seeking behavior (Spink, Ozmutlu, & Ozmutlu, 2002; Spink & Cole, 2005) while the latter has been concentrating on user behavior analysis (Shneiderman, 1978). Although the nodes representing these two authors are not closely with each other in Figure 2, they move nearer in the visualization of MFTACA method. However, the distance between Drs. Spink and McCain becomes farther in Figure 3 than Figure 2, indicating that our proposed MFTACA method separates authors sharing different research interests in knowledge domain maps. All of these facilitate the quality of maps in terms of the clustering performance.

**MDS-measurement**

In order to evaluate the performance of our proposed method quantitatively, we employ multi-dimensional scaling measurement (MDS-measurement) (Bu et al., 2016) to supplement our argument in “Network Science” section that in a quantitative way. Note that MDS-measurement is NOT the same as MDS. The basic principle of MDS-measurement is to calculate the MDS-measurement value ($\sigma$), which is equal to the ratio between the sum of the distance between the nodes within the same cluster ($c$), and the sum of the distance between the nodes in different clusters ($S$). Intuitively, a smaller $\sigma$ indicates a better clustering performance in knowledge domain maps in which nodes within the same cluster lie nearer while those in different clusters
lie farther. Table 3 shows the MDS-measurement result, where we can see that the MDS-measurement value ($\sigma$) of MFTACA is smaller than that of ACA, indicating a better clustering result in knowledge domain map. This confirms our observation in “Network Analysis” section.

Table 3. MDS-measurement results.

<table>
<thead>
<tr>
<th>Method</th>
<th>$c$</th>
<th>$S$</th>
<th>$\sigma (= c/S)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACA</td>
<td>546.61</td>
<td>4303.53</td>
<td>12.70%</td>
</tr>
<tr>
<td>MFTACA</td>
<td>513.17</td>
<td>4401.68</td>
<td>11.66%</td>
</tr>
</tbody>
</table>

Conclusions

This paper proposes a novel method combining the number of mentioned times, the number of context words, and the published time of references into traditional author co-citation analysis (ACA), called MFTACA (metadata-in-full-text author co-citation analysis). The results show that compared with traditional method, our newly proposed method not only shows better clustering performance but also provides more details in knowledge domain mappings. Considering that this method does not need a large volume of input such as content-based ACA (Jeong et al., 2014; Kim, Jeong, & Song, 2016), we believe that our proposed method could be easily applied to various disciplines so as to depict scientific intellectual structures by involving more information and improving traditional ACA.

Besides the method itself and its advantages compared with traditional ACA, this study provides several implications to the future researchers. First, we use full-text data but not intend to analyse content-level information, which breaks the conventional thinking to use complex natural language processing technologies to mine content- or semantic-level (Bu, Huang, Ding, & Ai, 2017) data so as to map knowledge domains. Second, such idea enables to be applied into not only ACA but also other scholarly network analyses, such as author bibliographic coupling analysis (Zhao & Strotmann, 2008a) and coauthorship analysis (Ding, 2011b). Last but not least, this research supplements the framework of bibliometric elements proposed by Morris and Vander Veer Martens (2008), in which papers, paper authors, paper journals, references, reference authors, reference journals, and index terms are included. Our study provides “citing sentences” as a bridge between “papers” and “references”, and shows the potential detailed affiliations about “citing sentences” such as the number of mentioned times and the number of context words of references. These have offered significant foundations for future supplements of the bibliometric element framework when more full-text data are involved.

However, there are several limitations in this research. For example, we only used first authors’ information instead of all authors’. The accuracy could thus be negatively affected. Moreover, there are still many other types of metadata that are not used to involve in ACA in previous studies and this paper, such as the sequence of co-cited authors (He, Ding, & Yan, 2012) and the number of figures or tables (Lee, West, & Howe, 2016). We would like to focus on these in the future.

References


Research on Multiple Technology Paths Extraction Method Based on Semantics: Take Magnetic Head of Hard Disk Drive as Example

Chen Liang¹ Yang Guancan¹ Shang Weijiao² Lei Xiaoping¹ Zhang Haichao¹

¹{chenl,yanggc,leixp,zhanghc}@istic.ac.cn
Center for Science and Technology Study, Institute of Science and Technology Information of China, Beijing

²shangwj490@126.com
Library of Chinese Academy of Forestry, Beijing

Abstract
On purpose of solving current limitation of patent citation analysis methods, we propose a new method named multiple technology topic paths method, which can capture more than one significant technology paths from patent citation network while patents from each path focus on certain technology topic. In more detail, this method uses text similarity between patents as weight of citation and applies dynamic programming for for technology paths searching in patent citation network, which gives rise to trajectories of main technologies development. As an empirical study, we take magnetic head of hard disk drive as example, the result shows the method can effectively generate whole picture of technological evolution process for magnetic head.

Conference Topic
Patent Analysis

Introduction
Technological evolution analysis is widely used within industry to support strategic decision and long-range planning. It provides access to exploring essential characteristics of technological evolution and examining relationships between different technologies, thus lies foundation for predicting future trends of emerging technologies.

Technological evolution analysis methods originated in 1940s, after development of almost 70 years, there are amount of members in this family, such as expert method, bibliometric methods, TRIZ (Russian abbreviation for Theory of Inventive Problem Solving), text mining method and so on. These methods help researchers to get insights into technological evolution from different viewpoints. However, at present all kinds of methods have their own limitations. Qualitative method such as the expert method is sensitive to subjective factors, quantitative methods can avoid this problem, but they have their own limitations, patent citation analysis is strongly dependent on citation network structure, data noise in network could lead to totally different technological evolution path, text mining method is proper for small dataset and TRIZ emphasizes its role in guiding specific innovation activity in engineering environment.

Under such condition, we propose a method to analyse technological evolution process in micro-level. This method applies text mining technique and dynamic programming algorithm on patent text and patent citation network for technological topic path searching. By extracting significant paths and combining them together, we can establish a network backbone which reflects the evolution process of significant technologies in target field.

In remainder of this paper, first we describe current development of technological evolution analysis methods based on patent citation as background knowledge in section 2. Second, the proposed method is explained in section 3. As an illustration, this method is applied to magnetic head of hard disk drive (HDD) in section 4. Finally, we conclude with a summary and briefly describe the further work.
Literature Review

Patent citation analysis

As the main outputs of R&D activities in companies, patents represent the direction and essential nature of technology development. Nowadays increasing researchers take patents as the main data source for technology management and intelligence analysis. Statistical data shows that 80% of the world's technological knowledge can be found in patent documents, furthermore, due to law procedure and commercial benefit, patent information is compiled rigorously under complete citation & audit mechanism, which not only leads to stronger correlation between citing patent and the cited one, but also the tendency in academic references has little impact on patent citations (Mitchell, 1991). Researchers have begun their efforts in patent citations since 1990s, while established a serial of indices based on patent citation, they proposed various methods for technological evolution analysis, within which the most representative one is so-called main path method. Compared to others, main path method evaluates one patent’s value by the structure of whole citation network and the position of this patent in it, which give rise to recognition for basic patents and milestone ones with high speed and low cost. Main path method was first proposed by Hummon and Doreian in 1989, this method takes importance inequality of citation relation into consideration, important citations act as main path for knowledge stream to go through, their remove will change the process of knowledge diffusion, while insignificant citation relations cause much less influence. In a sense, main path is the backbone of a citation network and it describes the most critical knowledge flow in it. Via main path method, researchers can focus on much fewer patents on main path, and achieve valuable result more easily. Main path method searches path by the following heuristic algorithm:

(i) Choose an index to measure significance of edges in citation network.
(ii) For each start point, pick the outward edge with maximal value among all outward edges from that start point. If more than one edge in the maximal value, take all them in.
(iii) Select the start points with maximal edge value among all start points in Step (ii). This is the start points of the main path.
(iv) Take the target nodes of the edges identified in the previous step. For each target node, pick the outward edges with maximal value from it. If these edges point to end points of the network, exit, else go back to Step (ii) and continue.

V. Batagelj (2003) first applied main path method on patent citation network with the background that patent citation network was one of few networks accessible at that time. Until 2007, B. Verspagen began to use main path method for technological evolution analysis, he noticed the original version of main path method uses greedy strategy for path searching, which can’t guarantee global optimum resolution. So he suggested a new version of main path method, which employs exhaustive strategy to ensure the result path to be global optimal. J. S. Liu called the original version of main path method as local main path method, and B. Verspagen’s version as global main path method, he suggested that it be useful to extract paths with top n weight by global path method, so more technological trajectories can be covered, he called this version as multiple Main Path method.

As a widely used approach, main path method has been integrated into popular network analysis software like Pajek. Although it provides efficient, and low-cost way to generate technology path, there are still some critical drawbacks in it which jeopardize its further application:

(1) Technology evolution is an uncertain process, it can develop in multiple possible ways and be implemented in various designs, these can be observed in patent citation network that multiple important technology paths occur simultaneously. Currently, both local main path
method and global main path method extract the “most important technology path” as output, which has drawback on technology coverage. Although B. Verspagen suggested extracting several technology paths to solve this problem, he did not implement it.
(2) Even technology path extracted by global main path method is more convincing than local main path, its time complexity is too high to tackle large-scale citation network.
(3) Indices based on network structure are sensitive to missing data and noise, which may cause significant change of result path.

**Methodology**

In order to solve main path method’s drawbacks, this paper takes topic into consideration and proposes a new variant of main path method which can extracts multiple main paths with each path focusing on a unique topic. So researchers can trace the evolution of different technologies conveniently. Furthermore, this method provides a way to observe relationship between different technologies. As patents in a main path belong to same topic, we named these new main paths as technology topic paths.

The whole process of this method is shown as below. First, searching patents and pre-processing them to build patent citation network. Second, taking patent abstracts as corpus and applying topic model to it to estimate model parameters, then calculating similarities between any pair of patent abstracts. Third, using proposed multiple main path method to generate candidate technology topic paths from patent citation network. Finally, sorting candidate technology topic paths by weight and generating technology topic backbone, the weakly connected components in technology topic backbone are the final technology topic paths.

**Data preparation**

After collecting relevant patents with their abstracts and citation information from patent database, data preparation includes two steps:

(1) Generating patent citation network by citation information;

(2) Pre-processing patent abstracts and building a corpus. Extract abstracts from patent dataset and run a series of pre-processing steps which include tokenizing, removing stop words, stemming, lemmatization and so on. Then use these cleaned patent abstracts to generate a corpus.

**Calculating semantic similarity between patent abstracts**

Apply topic model to the corpus to extract latent topics, and use these topics to represent every patent abstract. Compared with text of patent abstract, representation of patent abstract by topics can reduce number of features dramatically while maintaining necessary semantic information (Luo et.al. 2014). Furthermore, as patent abstract is short text, directly applying standard text similarity measure to patent abstracts measure always fails (Metzler, et.al. 2007) as a result of vocabulary mismatch problem (Furnas, et.al. 1987). However, since all topics occur in every patent abstract, patent abstract represented by topics can effectively avoid this problem. Here we choose the most widely used topic model--LDA (Blei, et.al. 2003) in our research, and calculate cosine similarity between different patent abstracts.

**Generation of candidate technology topic paths**

Intuitively speaking, main path with more significant topic should contain more patents and the semantic relationship among these patents should be tighter. Based on this, we create an indicator to measure path significance like this: calculating similarity of all possible patent pairs in a technology topic path, then summing all similarities as the weight of this main path. Main path weight is calculated by Eqs 1.
\[ W_{\text{path}} = \sum_{\text{node}_i \in \text{path}} \sum_{\text{node}_j \in \text{path}} \text{sim}(\text{node}_i, \text{node}_j) \]  

Where \( W_{\text{path}} \) indicates the weight of main path, and \( i, j \) are node IDs. \( \text{Sim}(\text{node}_i, \text{node}_j) \) indicates the semantic similarity between \( \text{node}_i \) and \( \text{node}_j \).

For example, there is a technology topic path shown as in Fig. 1, all node pairs on this path include (1,2), (1,3), (1,4), (2,3), (2,4), (3,4), therefore the weight of this path is: \( W_{\text{path}} = \text{sim}(1,2)+\text{sim}(1,3)+\text{sim}(1,4)+\text{sim}(2,3)+\text{sim}(2,4)+\text{sim}(3,4) \). Here we inverse the direction of citation to mark the direction of technology development.

**Figure 1. Technology Topic Path**

Given a patent citation network, the basic idea of candidate technology topic paths searching method is, start from a source node, traverse the citation network to find a path that has global optimal \( W_{\text{path}} \), in order to reduce this pathfinding algorithm’s time complexity, we use dynamic programming (Cormen, et. al. 2009) as search strategy.

In detail, candidate technology topic paths searching method is designed as below: taking a source node as start point and making breadth first traversal. Before cursor leaves current node, taking this node’s ID as key, taking a tuple, which consists of edge from start point to current node and this edge’s weight, as value and store the key-value pair into dictionary. After visiting all direct successors of source node, taking these successors as start points and repeat breadth first traversal. During traversal, if algorithm finds current node has been stored as key in dictionary before, it means there exists another path from the source node to current node that cursor has already traversed, then the algorithm will achieve this earlier path and compare its \( W_{\text{path}} \) with \( W_{\text{path}} \) of current path, if current path’s \( W_{\text{path}} \) is greater, the algorithm will create a tuple consisting of current path and its \( W_{\text{path}} \), then use it to replace the value corresponding to current node in dictionary, else keep the value unchanged and continue traversal process until the end. At last dictionary stores paths from the source node to all end nodes it can achieve, and every path is the global optimal one among all possible paths this path’s both end. We output path with largest \( W_{\text{path}} \) as result.

![Diagram of Technology Topic Path](image-url)
Figure 2. Patent Citation Network

In order to make the searching process clear, let’s assume that we have a patent citation network shown in Fig. 2, the semantic similarities between patent abstracts are shown in Table 1, here we show how to search candidate technology topic path in Fig. 3, the grid in each step presents dictionary’s content at that moment.

Table 1. Semantic Similarities Between Patents

<table>
<thead>
<tr>
<th>Node ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>0.4</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
<td>1.0</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>0.5</td>
<td>1.0</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>4</td>
<td>0.3</td>
<td>0.3</td>
<td>0.4</td>
<td>1.0</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>5</td>
<td>0.4</td>
<td>0.2</td>
<td>0.3</td>
<td>0.5</td>
<td>1.0</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>6</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.6</td>
<td>1.0</td>
<td>0.4</td>
</tr>
<tr>
<td>7</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>0.4</td>
<td>0.4</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Figure 3. Technology Topic Path Searching Process
In step (1), from node 1 we begin breadth first traversal and achieve its successor nodes 2, 3, then we put nodes 2, 3 and their relevant information into dictionary. For example, we take node 2 as key, take tuple consisting of edge from node 1 to node 2 and its weight as value, and stores the key-value pair in dictionary; In step (2), we visit node 2’s successor node 4, according to Eqs 1, we get weight of path $1 \rightarrow 2 \rightarrow 4$, then take node 4 as key, take tuple including path $1 \rightarrow 2 \rightarrow 4$ and its weight as value and put them into dictionary; In step (3), we get node 3’s successors node 4, 5, since node 4 has been put into dictionary as key before, we need to calculate the weight of new path from node 1 to node 4, which is $1 \rightarrow 3 \rightarrow 4$, then compare it with the weight of earlier path, which is $1 \rightarrow 2 \rightarrow 4$, as earlier path’s weight is greater than new path’s, we keep node 4’s corresponding value unchanged in dictionary, but if earlier path’s weight is less than new path’s, we should replace node 4’s corresponding value with tuple consisting of path $1 \rightarrow 3 \rightarrow 4$ and its weight, then take node 5 as key, take tuple contains path $1 \rightarrow 3 \rightarrow 5$ and its weight as value, then add them into dictionary, and so on in step (4), (5). Eventually we find $1 \rightarrow 3 \rightarrow 5 \rightarrow 6$ as the path with highest weight and return it as the candidate technology topic path.

Generation of technology topic backbone

On basis of candidate technology topic paths, now we generate technology topic backbone. The process of generating technology topic backbone is, first we sort all candidate technology topic paths by weight, then we take the path with highest value as target network, and merge other topic paths into target networks in descendent order. After every mergence, record the snapshot of current target network, then we select a snapshot to be the final technology topic backbone. The intuition of selecting strategy is that technology topic backbone should highlight a few important technology topic paths, which are indicated by path’s weight, in the meanwhile, as paths merge into target network, paths with different topics may be connected, thus we have to balance number of highly weighed topic paths against number of weakly connected components in target network, and make an optimal choice from snapshots as technology topic backbone.

Illustrative example: Technology topic paths in HDD

The hard disk drive (HDD) technology field is selected as an exemplary study, because changes in technology of HDD have been much rapider than that in other fields, so it provides an opportunity of technological evolution analysis in a shorter time (Christensen, 2011). America is the birthplace of HDD and it keeps playing an essential role in propelling HDD technology move forward, major HDD companies have paid much attention to intelligence protection and filed amount of patent applications to protect their intelligence right since the beginning. Therefore relevant patents can be treated as a more representative and trustworthy data source for technological evolution analysis. Since the first appearance of HDD in 1956, HDD has undergone 3 generations which includes ferrite head disk drive, thin film head disk drive and magnetoresistant head disk drive (Christensen. 2011). As thin film head disk drive has endured the longest period of development, here we focus on thin film head which is a core component of thin film head disk, and collect bibliographic and textual information of patents on HDD from USPTO database.

Retrieving Patents

(1) Take (TTL//"thin film head" OR ABST//"thin film head") AND APD//1/1/1976->31/12/2013 as search query and run it on USPTO official website, we achieve 190 patents.
(2) With the assistance of ISTIC Patent Database supported by Institute of Science and Technology Information of China (data updated til 38th week in 2014), we retrieve patents
citing or cited by 190 patents. According to these patents’ USPC labels, we removes irrelevant ones and generate a dataset of 3503 patents.

(3) With citation information of 3503 patents, we generate a citation network consisting of 15 weakly connected components, among which the giant connected component contains 3193 patents. After removing invalidated patents, there are 2876 patents left. We use bibliographic and textual information of these 2876 patents for our exemplary study. As for patents in other weakly connected components, since their isolation, ignorance of them induces little negative effect on analysis result.

Calculating semantic similarities between patent abstracts

We use Python and NLP (Natural Language Processing) package Gensim (Redim, et.al. 2010), NLTK (Bird, et. al. 2009) to process patent abstracts. NLTK provides stop words list and term stemmer, Gensim provides tf-idf function, LDA model function and Cosine function to calculate similarities between patent abstracts.

After pre-processing patent abstracts as shown in previous section, we build a corpus and a vocabulary consisting of 4707 terms. According to perplexity curve generated by the corpus, we set topic number to be 110 and estimate parameters using LDA model. Finally, with representation of every patent abstract by these 110 topics, we calculate semantic similarities for every possible pair of patents.

Generating candidate technology topic paths

We implement technology topic path extraction algorithm using Python and package Networkx (Hagberg, et.al. 2008) which provides a series of basic network operation functions. After running technology topic path extraction algorithm on patent citation network, we get 828 candidate technology topic paths. A more detailed information of these paths is shown in Fig. 4, where the horizontal axis represents number of nodes contained in technology topic path, the vertical axis represents number of paths with same number of nodes. In order to show the changes of path counts as nodes of path increases, we smooth histogram by Gaussian kernel density estimation (KDE) and generate a blue line. From Fig. 4, we can see that most of the paths are concentrated in interval [4,9] with a few ones scattering at both ends of the curve.

![Figure 4. Distribution of Nodes in Technology Topic Paths](image-url)
Then we analyse the distribution of paths’ weight, due to the fact that path weight is calculated based on semantic similarity, which causes a fact that any two paths’ weight are seldom equal to each other, so we use KDE to smooth path weight distribution as before, and the result is shown in Fig. 5, where the horizontal axis represents path weight and the vertical axis represents number of paths falling in a weight interval. From Fig. 5 we can see most paths’ weight exist between interval (0,8]. As path weight increases, the number of paths shows a descendent trend.

Figure 5. Distribution of Weight for Technology Topic Paths

Generation of technology topic backbone

Based on 7 distinguished technology topic paths recognized in Figure 5, we establish technology topic backbone as Figure 6. We assign nodes on different path with different colour, as to some paths are crossed at one node, we assign any colour to the node from these paths. The labels besides node are patents number, and we use edges width to indicate texts similarity between patents.

Figure 6. Technology Topic Backbone
Result analysis

Detailed information about the 7 final technology topic path are shown in Table 2. We can see from 1972 to 2011, innovation activities of thin film head mainly involve 7 topics. In fact, according to relevant domain knowledge, major performance indices of HDD include areal density, absolute output, signal-noise-rate and resistance (Keiichi, et.al. 2006). The technology topic backbone indicates that innovators advanced HDD performance from two aspects: improvement of specific components and modification of the whole structure of technology. Path 2— innovations for layer of thin film head, path 4— innovations of read-write component for thin film head, path 7— innovations for magnetic core in thin film head belong to the first aspect, and Structure design of thin film head belongs to the second one.

Table 2. Information of Main technology topic paths

<table>
<thead>
<tr>
<th>Path ID</th>
<th>Nodes number</th>
<th>Technology topic</th>
<th>Last time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>Structure design of thin film head</td>
<td>1980-2010</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>Manufacture technology for layer of thin film head</td>
<td>1986-2009</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>Manufacture technology for thin film head with magnetoresistant components</td>
<td>1977-2004</td>
</tr>
<tr>
<td>4</td>
<td>36</td>
<td>Manufacture technology of read-write component for thin film head</td>
<td>1972-2011</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>Etching method in thin film head manufacture</td>
<td>1976-2006</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>Innovation of magnetoresistant head</td>
<td>1976-2010</td>
</tr>
<tr>
<td>7</td>
<td>11</td>
<td>Innovations for magnetic core in thin film head</td>
<td>1982-2005</td>
</tr>
</tbody>
</table>

Besides innovations focusing on advancing performance for HDD, other significant development of thin film head technologies includes path 5—etching methods in thin film head manufacture, which are used in manufacture process of thin film head. Also we can identify the tight relationship between thin film head and magnetoresistant head from path 3 and path 6. Actually the purpose for thin film head to integrate magnetoresistant components is to create a separate reading head, which enable to read very small magnetic features reliably (Mao. 2013). As to why related technologies of thin film head, such as etching method in thin film head manufacture and manufacture technology for layer of magnetic head, can keep on developing when magnetoresistant head has occupied dominant position in the market, the answer is that, as a result of structural inheritance between thin film head and magnetoresistant head, these technologies relevant to thin film head can be applied on magnetoresistant head and facilitate its development.

Conclusion and discussion

This paper proposes a new main path analysis method which takes topic into consideration comparing to previous methods of technological evolution analysis, it can produce multiple main paths each of which focuses on a unique technology topic. Thus apart from identifying major directions of innovation in a technology field, new method enable researchers to trace each direction’s evolutionary trajectory at micro level, in the meanwhile, with all multiple main paths put together, researchers have the opportunity to explore relationship between
different technology topics and get a deep understanding of technological evolution in a field. In addition, this method is domain-independent, in another research project we applied it to field of electric vehicle battery and precisely identified two most significant technology topics which includes cooling technology for battery module and monitoring technology for electric power of battery.

There are a lot of improvement for our method, such as the indicator used to measure significance of technology topic path. It’s probable that high value is caused by path with a large number of nodes in it, and it doesn’t necessarily reflect significance of this path. Vice versa, for some emerging technologies with tremendous potential, they may not be highlighted by path weight indicator as a result of a few patents involved. So in next step, we will focus on measuring the potential of emerging technologies.

References


Relationships between Venture Capital and Listed Firms’ Technology Innovation: In the View of Patent, Growth Stage, and Industries

Guo Ying1  Zhang Lili2  Song Haijie3  Huang Danyang4  Meng Wei5  Liu Xiaoyu6

1 guoying_bit@163.com
Beijing Institute of Technology, Beijing (China)

2 zhanglili_bit@163.com
Beijing Institute of Technology, Beijing (China)

3 754478689@qq.com
Beijing Institute of Technology, Beijing (China)

4 hjsong@yeah.net
Beijing Institute of Technology, Beijing (China)

5 mengwei199132@bjtu.edu.cn
Beijing Jiaotong University, Beijing (China)

6 xiaoyu.liu2019@foxmail.com
Beijing Institute of Technology, Beijing (China)

Abstract
Patents signal a startup’s technological and innovative capability. Venture Capital (VC) investors can evaluate the true potential and behaviour of startups through their patents. While, there are many factors might have effect on the amount of VC financing. Some scholars have researched this question in other countries, but the detailed interaction between VC and patent activities in China is different and attractive. This paper tries to explore this relationship in the view of patent, growth stage, and industries by hypothetical test. We first propose four hypotheses about influence factors to VC. Then we use multi-source to get patent and VC data of listed companies. By descriptive statistics, univariate analysis and multivariate analysis, we try to verify our hypotheses. This research can provide advices to financers, investors, and policy makers by describing the relationship between technology innovation and VC.

Keywords: Venture Capital; Patent Analysis; Technology Innovation; China; Hypothetical Test

Conference Topic
Indicators; Patent analysis; Innovation and creativity

Introduction
Innovation and its role as crucial carrier of sustainable national economic progress and competitive edge has been acknowledged to certain extent in formal writings since Schumpeter (Lazzarotti, Dalfovo, Hoffmann & Valmir, 2011). Many scholars argue that cost input is good for nurturing innovation (Gans, Hsu & Stern, 2002; Lerner, 2012; Chemmanur, Loutskina & Tian, 2014). However, some firms, especially new firms, can rarely rely on internal cash flows in their pursuit of innovation. Among the sources of external finance which may help to entrepreneurs, capital funding can provide not only the financial resources they require, but also assistance to enhance the design, development, and performance of portfolio companies (Schwienbacher, 2008; Cumming & Schwienbacher, 2005; Luukkonen, Deschryvere & Bertoni, 2013). Investment decisions largely characterised by the management thinking and quality of desired goals and objectives according to the current development state (Zacharakis & Meyer, 2000). For a commercial enterprise, quality is a state that can be reflected by employee’s skills, product or service designing for target customers, supply chain setup, proprietary rights, budgets, sale, income, marketing research and activities (Teece, 1993; Tripsas, 1997). Unlike
established firms that own many of these assets, new firms face a particularly strong challenge in securing complementary assets for their technology commercialization. Therefore, venture capital (VC) emerged and gives new firms more opportunities. After the world’s first VC firm being founded in 1946 in the US, VC begins its mature organization and institutionalization stage. This industry has changed the way of technological innovation in the US, which greatly promoted the sustained national economic growth (Yang, Wang & Wen-Jie, 2016). As to China, VC started in the 1990s and has gradually embarked on the development track after those years (Long & Liu, 2013). However, the relationship between new firms and VCs is characterized by information asymmetries because it is difficult for VCs to evaluate the true potential and behaviour of new firms (Baum and Silverman, 2004; Neher, 1999). Also, it is hard to estimate how the VC would influence on enterprises’ innovation output.

To give investors more significant information, some technology-based firms may utilize patents to communicate the quality of their underlying technologies to investors (Hsu and Ziedonis, 2013; Conti et al., 2013; Haeussler et al., 2014). Patent is an important output of firm innovation, and would lead to competitive advantage due to asset specificity. Patents signal a startup’s technological and innovative capability and have direct effects on VC financing. On the other side, many variables, such as VC funding, have been investigated that they have influence on patents’ effectiveness. Previous studies provide plenty theoretical explanation and empirical test through resource based view on these topics.

However, the exploration between VC and patents activity is few in China. China has a notable patents number in recent years, especially after the national strategy of “building innovation-oriented country” in 2006, and remains the biggest patent application country since 2011. The Derwent World Patents Index 3 estimated that by 2015, Chinese owned companies would be generating 500,000 patents, a number 25% more than the number of patents the US is expected to have. Even though, China’s market environment is different from that of other countries, how the VC and new firms’ technology innovation interacts each other is still not clear. Therefore, this paper tries to analyse their interaction in the view of different industries, bio-pharmaceutical and electrical & electronic. There are numbers of differences among industries, including government’s attitude and policy, global developing environments, the importance of technology innovation and investors’ opinion. To understand the different interactions of different industries’ process, we test several hypotheses using a sample of bio-pharmaceutical and electrical & electronics companies that seek VC. We analysis the VC data from PEDaily.cn [http://zdb.pedaily.cn/] and patent data from SIPO. Next section will introduce some theories and our hypotheses, then introduce the data and the method to test hypothesis, and results follow. Conclusions and discussion are in the last part.

**Theory and hypotheses**

*Technology innovation and patent*

Innovation is defined as the creation of new knowledge and ideas to facilitate business outcomes, processes and structures and creates new products and services (Du Plessis, 2007). Technology innovation refers to be a kind of improvements and modifications of existing technologies, and creating of new technologies. It is driven by the creation of knowledge and knowledge creation is perceived as one of major assets (Pei, 2008). Several measurement indicators of technology innovation have been proposed (Pakes and Griliches, 1984; Tseng, 2009; Coccia, 2014; Ayerbe, Lazaric, Callois & Mitkova 2014), including R&D, patents and new products. As a description of the fundamental nature of technology, or knowledge, and the strategic behavior of the firms when utilizing these technologies (Woo, Jang, & Kim 2015), patent is used as a major measure to elucidate technology innovation in this study for the following reasons (Tseng, 2014). First, patent data are easily accessible via the database, and don’t have problems of imprecise
definition and lack of comparability between firms. Second, patent data are accurately recorded and easily manipulated for more detailed analysis by technological field, patent inventors and assignees. Third, patents are an objective measure of technology innovation, since patents are examined and eventually granted by a single national patent office. Finally, in comparison with other sources, patents are often the only timely measure of rapid technological change, particularly in the context of global competition.

To judge whether there are relationships between technology innovation and VC, we propose the two hypotheses:

**Hypothesis 1.** Ceteris paribus, whether a startup has patents or not will make significant sense on the amount of funding money obtained from VCs.

**Hypothesis 2.** Ceteris paribus, the more patents a startup holds, the more funding money will be obtained from VCs.

*The growth stage of financing firms*

A study on China has proved that firm ageing has positive impact on organisation technology innovation outputs (Shi, Zhu 2014). To simplify our analysis, we consider the growth theory, instead of every-year age, to check the relationship between firm’s growth stage and its financing ability. Growth theory suggests that appropriate growth and capital capacity strategies vary at different stages of a firm’s growth (Anthony & Ramesh 1992), which consists of four stages: in the early, development, expansion and profit stage. Each stage exhibits significant differences in terms of situation, organizational strategy, structure, and decision-making style (Adizes, 2004; Miller and Friesen, 1984; Pashley and Philippatos, 1990). So, we get information of growth stage from our database and use 1~4 to respectively represent four stages: in the early, development, expansion and profit stage. So we propose this hypothesis:

**Hypothesis 3.** There are significant variations between the stage of growth and the amount of VC financing.

*Characters of different industries*

Every industry has its own character both on internal technology and external environment. In the view of internal technology, numerous complementary technologies are developed via different and diverse R&D lines in a complex industry, which increase the overall probability of creating a successful innovation (Bessen and Maskin, 2009). So, different industries have different attitudes to the relationship between technology innovation and patent, including the “innovative degree” contained by each patent in different industry, the importance of technology innovation to Return on Investment (RoI) and enterprise performance, which has direct influence on investors’ decisions. As to external environment, government usually holds different attitude and carries out various policies to different industries. At the same time, the maturity of development varies from industry to industry, which means their global marketing environments are also different. All these factors will work on investors’ judgement. Therefore, we decide to compare two typical and important industries, pharmaceutical industry and energy exploitation industry. Our second hypothesis is:

**Hypothesis 4.** In different industries, the number of startup’s granted patents, the growth stages of financing firms, investors’ funding amount, and VC series will have different characteristics.

*Data and sample*

*Data and sample description*

We study the role of information generated by corporation, especially patent, for VC financing in China. In this study, we choose listed companies as object of study. Because they will make more information (financial performance, corporation governance, the amount of VC funding,
and so on) public and it’s too difficult to find enough information of unlisted companies to analyse. The VC data is mainly collected from China Venture (https://www.chinaventure.com.cn/) and complementally from PEdata (http://www.pedata.cn/), both these two databases are relatively complete, widely used and have better reputations. The patent data is collected from SIPO (State Intellectual Property Office of the P.R.C., http://www.sipo.gov.cn/) and CSMAR (http://www.gtafe.com/), a professional database for research in China. Here we distinguish granted patent and applied patent (Haeussler, Harhoff, & Mueller 2009), cause only applied patent can’t evaluate technology innovation of a firm, while granted patent by SIPO can provide more reliable information about technology innovation.

**Measures**

There are two dependent variables in our analysis, whether obtained VC or not (VC_obtain) and the amount of VC funding (VC_amount). What’s more, invention patent, utility patent, and design patent have different characters and different degree of innovation. So, we will also distinguish the patent’s type and use them as independent variables. According to previous study (Hoenen, Kolympiris, Schoenmakers, & Kalaitzandonakes 2014), we firstly conduct univariate analysis by using whether applying and being granted patent (3 types) or not, the number of applied patents (3 types), the number of granted patents (3 types) to analyse whether these factors will have effect on getting financing or not.

**Table 1. Variables and their implication**

<table>
<thead>
<tr>
<th>Variable’s kind</th>
<th>Variable’s name</th>
<th>Implication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent variable</td>
<td>VC_obtain</td>
<td>Whether a financing firm obtained VC or not, 1 for yes and 0 for no.</td>
</tr>
<tr>
<td></td>
<td>VC_amount</td>
<td>The amount of VC funding a firm get in one VC event, counted by one million RMB.</td>
</tr>
<tr>
<td></td>
<td>Apply</td>
<td>Whether a financing firm have applied patent or not, 1 for yes and 0 for no.</td>
</tr>
<tr>
<td></td>
<td>Grant</td>
<td>Whether a financing firm have been granted patent or not, 1 for yes and 0 for no.</td>
</tr>
<tr>
<td>Independent variable</td>
<td>Patent</td>
<td>The whole number of applied patents that a financing firm held.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The whole number of granted patents that a financing firm held.</td>
</tr>
<tr>
<td></td>
<td>Invention Patent</td>
<td>INVapply</td>
</tr>
<tr>
<td></td>
<td></td>
<td>INVgrant</td>
</tr>
<tr>
<td></td>
<td>Utility Patent</td>
<td>UTIapply</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UTIgrant</td>
</tr>
<tr>
<td></td>
<td>Design Patent</td>
<td>DESapply</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DESgrant</td>
</tr>
</tbody>
</table>
Stage of lifecycle of financing firm, use 1~4 to respectively represent four stages: in the early, development, expansion and profit stage. The industry of financing firm.

Secondly, we will further add the stage of lifecycle of financing firm as independent variable to analyse how these variables operate on VC_amount. Table 1 describes these variables in detail. In this step, we will build a model to evaluate the influence of these variables on VC_amount by empirical test. We propose the model as follows:

\[ VC_{\text{amount}} = \alpha_0 + \alpha_1 \text{PATapply}_i + \alpha_2 \text{PATgrant}_i + \alpha_3 \text{INVapply}_i + \alpha_4 \text{INVgrant}_i + \alpha_5 \text{UTIapply}_i + \alpha_6 \text{UTIgrant}_i + \alpha_7 \text{DESapply}_i + \alpha_8 \text{DESgrant}_i + \alpha_9 \text{Stage} \]

Every \( \alpha \) is a regression coefficient of each independent variable. If \( \alpha \) is positive, it means its corresponding variable has positive effect on obtaining more financing amount.

Thirdly, we consider the function of control variable, the industry of financing firm (Industry). There are many studies, as well as foregoing analysis have proved the relationship between patent and VC is only obvious in A round (Conti, Thursby, & Thursby 2013; Zhou, Sandner, Martinelli, & Block 2016), so we choose all data about A round as a sub-database to continue our analysis. According to our now available database, fewer companies are included in each industry (provided by CSMAR), so we cleared up our data of A round, clustered and chose two industries, bio-pharmaceutical industry and electrical & electronic industry, to conduct later research.

Bio-pharmaceutical industry refers to economic activities of mass production of marketable drugs through the application in pharmaceutical industry of the research results of modern biotechnologies, such as genetic engineering, cell engineering, enzyme engineering, fermentation engineering and protein engineering. In many countries, bio-pharmaceutical industry has been identified as a key industry in the 21st century. In comparison, the bio-pharmaceutical industry is characterized with high-tech, high investment, high-risk, high-yield and long-cycle.

Intellectual property rights are essential to innovation and development of the bio-pharmaceutical industry. Only by establishing a sound intellectual property system, can the development of bio-pharmaceutical industry form the virtuous circle of “innovation-protection-innovation again”. Due to the lack of protection consciousness and talents, protection on intellectual property related with biomedical field is inadequate in China. Both the application of and the protection on patents of biomedical research achievements have a big gap with foreign countries, which result in some high value research results failed to properly obtain patent protection. Moreover, when cooperating with industrial companies, China’s biomedical research institutions often cannot accurately assess the value of their research achievements.

Analysis and results

Descriptive statistics

We searched VC and patent date from January 1st, 1999 to December 31st, 2016. After integrating the same type of data from different sources, matching VC records and patent records of the same corporations, cleaning fuzzy records, we finally got 19,115 records of companies’ applying patents and 2,757 events of VC of listed companies. It’s worthy to note that, in the view of corporations who obtained VC, the proportion of having patent is 15.05% and 20.38% of having no patent, which doesn’t have much special meanings cause the database of applying patent is relatively bigger than that of VC. The detail of all these data is shown in Figure 1. There are some corporations who have financed more than once, to get more valuable
samples, we use one event of VC as a sample. So, we get 4690 samples, including 2,757 VC events and 1,993 corporations who have patent but no VC funding.

1,993 corporations who have patent but no VC funding.
(18,577 records of corporations who have patent but no VC funding.)

353 corporations who have patent and VC funding.
(538 records of corporations who have patent and VC funding.)

1,732 corporations, 2,757 VC events
2,346 corporations, 19,115 records of applying patents

1,732 corporations, 2,757 VC events
2,346 corporations, 19,115 records of applying patents

1,379 corporations who have no patent but VC funding.
(2,219 records of corporations who have no patent but VC funding.)

Figure 1. Distribution of all VC data and patent data

Table 2 shows the result of descriptive statistics of main variables. The standard deviations of both the amount of VC funding and the number of patent are large, which illustrates that every firm’s technology innovation varies and the results of their VC financing are also significantly different. Among, Semiconductor Manufacturing International Corporation (SMIC) financed the most in one event, 7020.91 million RMB from Shenzhen Capital Group (SCGC), China Merchants & Fortune Assets Management Ltd, and Zhangjiang Hi-Tech Park in its A round financing. And ZTE Corporation has applied and been granted the most patents, 40,608 and 20,355 respectively. What’s more, a large number of top patent holders, including ZTE, Sinopec, Gree, PetroChina, BYD, Midea, JAC, Baosteel and many “successful” companies, have never financed for VC before, this might because they have good market performance and reputation to get enough capital flow. This also explains why the relationship between VC and patent is more obvious in A round and growth-oriented small and medium enterprises.

Table 2. Descriptive statistics of main variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Sample size</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>VC_obtain</td>
<td>4690</td>
<td>0.580</td>
<td>0.494</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>VC_amount</td>
<td>2757</td>
<td>26.494</td>
<td>145.347</td>
<td>0</td>
<td>0.023</td>
<td>7020.91</td>
</tr>
<tr>
<td>Apply</td>
<td>2346</td>
<td>0.533</td>
<td>0.499</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Grant</td>
<td>2346</td>
<td>0.503</td>
<td>0.500</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PATapply</td>
<td>4690</td>
<td>133.027</td>
<td>948.377</td>
<td>12</td>
<td>0</td>
<td>40608</td>
</tr>
<tr>
<td>PATgrant</td>
<td>4690</td>
<td>83.093</td>
<td>552.289</td>
<td>1</td>
<td>0</td>
<td>20355</td>
</tr>
<tr>
<td>INVapply</td>
<td>4690</td>
<td>59.696</td>
<td>700.101</td>
<td>1</td>
<td>0</td>
<td>37253</td>
</tr>
<tr>
<td>INVgrant</td>
<td>4690</td>
<td>20.480</td>
<td>306.341</td>
<td>0</td>
<td>0</td>
<td>17094</td>
</tr>
<tr>
<td>UTapply</td>
<td>4690</td>
<td>53.158</td>
<td>297.040</td>
<td>0</td>
<td>0</td>
<td>7644</td>
</tr>
<tr>
<td>UTgrant</td>
<td>4690</td>
<td>45.977</td>
<td>277.363</td>
<td>0</td>
<td>0</td>
<td>7240</td>
</tr>
<tr>
<td>DESapply</td>
<td>4690</td>
<td>20.173</td>
<td>120.411</td>
<td>0</td>
<td>0</td>
<td>3461</td>
</tr>
<tr>
<td>DESgrant</td>
<td>4690</td>
<td>16.635</td>
<td>109.133</td>
<td>0</td>
<td>0</td>
<td>3388</td>
</tr>
</tbody>
</table>

We finally get 538 records of corporations who have patent and VC funding. These records come from 353 listed corporations, including 5 types of financing rounds (as shown in Figure 2) 54 industries (top 10 industries are shown in Figure 3).
Among the whole 538 records (Figure 3), investments in mechanical equipment, hardware industry, communication equipment and other high-tech related industries dominate. On account that high-tech development is an exit to powerful nation, China has turned into the age of technology innovation. Besides, these technologies bring so much benefits to humans, which makes our life more convenient and fantastic. Manufacture is a big industry, which is closed to daily life. For example, we need furniture, vehicles, and other necessities. So, this industry also attracts so many investors, and industry pressure is not too small. The third one is pharmaceutical industry. This is a traditional, fast developing industry, for the reason that people are getting more concerned on health. Many new medicines will be created, even may change the word.

**Figure 2. Distribution of financing rounds**

**Figure 3. Distribution of top 10 financing industries**

*Univariate analysis*

To initially check whether a financing firm having patent (both applied and granted) will or not have influence on VCs’ decision (investing or not), we use univariate analysis to reach this goal. We divide all our data about 2,757 listed company into two group, have ever got VC, and never got VC. Then we use T test for independent-samples and the result is shown in Table 3. The Wilcoxon rank-sum test also shows there are differences among these variables at 0.01 significance level. Comparing to corporations who have never got VC, the corporations who have ever got VC usually have larger numbers of both proportions and amount of applying patents and granted patents, except the design patent. This illustrates that those corporations who have applied patents and been granted patents tend to obtain VC more easily from investors, as well as there is indeed close relationship between VC patent, which verifies Hypothesis 1. Therefore, we will further analyse how these variables work on the amount of VC financing by the method of multivariate analysis.
Table 3. Univariate analysis

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>T test for independent-samples (t score)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Have ever got VC</td>
<td>Have never got VC</td>
</tr>
<tr>
<td>Apply</td>
<td>1.00</td>
<td>0.20</td>
</tr>
<tr>
<td>Grant</td>
<td>0.94</td>
<td>0.19</td>
</tr>
<tr>
<td>PATapply</td>
<td>254.01</td>
<td>45.48</td>
</tr>
<tr>
<td>PATgrant</td>
<td>162.81</td>
<td>25.41</td>
</tr>
<tr>
<td>INVapply</td>
<td>110.20</td>
<td>23.15</td>
</tr>
<tr>
<td>INVgrant</td>
<td>40.04</td>
<td>6.32</td>
</tr>
<tr>
<td>UTIapply</td>
<td>102.58</td>
<td>17.39</td>
</tr>
<tr>
<td>UTIgrant</td>
<td>88.83</td>
<td>14.97</td>
</tr>
<tr>
<td>DESapply</td>
<td>41.24</td>
<td>4.93</td>
</tr>
<tr>
<td>DESgrant</td>
<td>33.94</td>
<td>4.11</td>
</tr>
</tbody>
</table>

***. significantly correlated at 0.01 level  
**. significantly correlated at 0.05 level  
*. significantly correlated at 0.1 level

Multivariate analysis

Since we have verified that who have applied patents and been granted patents tend to obtain VC more easily from investors, we will only analyse the data of 538 records of corporations who have both granted patent and VC funding.

Table 4. Multivariate analysis

<table>
<thead>
<tr>
<th>Variables</th>
<th>VC_amount</th>
<th>PATapply</th>
<th>PATgrant</th>
<th>INVapply</th>
<th>INVgrant</th>
<th>UTIapply</th>
<th>UTIgrant</th>
<th>DESapply</th>
<th>DESgrant</th>
</tr>
</thead>
<tbody>
<tr>
<td>VC_amount</td>
<td>1.00</td>
<td>-.028</td>
<td>-.028</td>
<td>-.024</td>
<td>-.025</td>
<td>-.029</td>
<td>-.026</td>
<td>-.025</td>
<td>-.019</td>
</tr>
<tr>
<td>PATapply</td>
<td>-.028</td>
<td>1.000</td>
<td>.969**</td>
<td>.976**</td>
<td>.976**</td>
<td>.941**</td>
<td>.953**</td>
<td>.337**</td>
<td>.352**</td>
</tr>
<tr>
<td>PATgrant</td>
<td>-.028</td>
<td>.969**</td>
<td>1.000</td>
<td>.896**</td>
<td>.919**</td>
<td>.972**</td>
<td>.982**</td>
<td>.487**</td>
<td>.510**</td>
</tr>
<tr>
<td>INVapply</td>
<td>-.024</td>
<td>.976**</td>
<td>.896**</td>
<td>1.000</td>
<td>.980**</td>
<td>.860**</td>
<td>.883**</td>
<td>.182**</td>
<td>.198**</td>
</tr>
<tr>
<td>INVgrant</td>
<td>-.025</td>
<td>.976**</td>
<td>.919**</td>
<td>.980**</td>
<td>1.000</td>
<td>.891**</td>
<td>.898**</td>
<td>.224**</td>
<td>.239**</td>
</tr>
<tr>
<td>UTIapply</td>
<td>-.029</td>
<td>.941**</td>
<td>.972**</td>
<td>.860**</td>
<td>.891**</td>
<td>1.000</td>
<td>.988**</td>
<td>.368**</td>
<td>.384**</td>
</tr>
<tr>
<td>UTIgrant</td>
<td>-.026</td>
<td>.953**</td>
<td>.982**</td>
<td>.883**</td>
<td>.898**</td>
<td>.988**</td>
<td>1.000</td>
<td>.366**</td>
<td>.391**</td>
</tr>
<tr>
<td>DESapply</td>
<td>-.025</td>
<td>.337**</td>
<td>.487**</td>
<td>.182**</td>
<td>.224**</td>
<td>.368**</td>
<td>.366**</td>
<td>1.000</td>
<td>.988**</td>
</tr>
<tr>
<td>DESgrant</td>
<td>-.019</td>
<td>.352**</td>
<td>.510**</td>
<td>.198**</td>
<td>.239**</td>
<td>.384**</td>
<td>.391**</td>
<td>.988**</td>
<td>1.000</td>
</tr>
</tbody>
</table>

We can see that the VC_amount doesn’t have correlations with PATapply, PATgrant, INVapply, INVgrant, UTIapply, UTIgrant, DESapply or DESgrant. These results reject Hypothesis 2, which is opposite to previous study. By analysing our data and China’s situation, we find that all the patent data we got is about listed companies. While it is a signal that the company already has good performance, they can get enough capital flow through other ways, not only VC. What’s more, VC still doesn’t develop maturely in China and many investors tend to make their decisions without considering patents.

Other analysis

To test Hypothesis 3, we calculate the average of financing amounts of different growth stages of financing firms. The 538 records of corporations who have patent and VC funding are analysed and the result is shown as Table 5. We can see the averages are different and the amount is larger in earlier stage. This because venture capitalists are primarily engaged in nurturing innovation in small, nimble startups (Morgan, 2014). Therefore, a firm in early stage tends to get VC easily, which verifies Hypothesis 2.
As for different industries, pharmaceutical industry and energy exploitation industry, we analyze some dimensionalities and get the result shown in Table 6. We can see that pharmaceutical industry get less VC funding with more patents, which indicates pharmaceutical industry is more technology-oriented than energy exploitation industry. And this conclusion is also in agreement with the common sense that pharmaceutical industry will have more patents when creating new medicines. The distributions of growth stages of financing firms in these two industries don’t make big difference and both are focus on development stage and expansion stage. This might because these two industries have much risks in early stage and VC investors prefer other industries for risk averse, which also result that they have more series A in their development stage and expansion stage instead of early stage. While coming to these two stages means they have a promising market and they need more funding for further
expansion, in which time VC will play an important role. Therefore, the results verify Hypothesis 4 partly so we need further study in the future.

Conclusion
This paper first described the relationship of whole VC data and patent data, distributions of financing rounds and industries. Among all the corporations who obtained VC, the proportion of having patent is 15.05%. The mainly financing round is A round and high-tech related industries dominate. We put forward four hypotheses based on insights from earlier research. Hypothesis 1 is whether a startup has patents or not will make significant sense on the amount of funding money obtained from VCs, which is verified through the univariate analysis. Hypothesis 2 is the more patents a startup holds, the more funding money will be obtained from VCs. However, this can’t be verified which might because Chinese characteristic condition or our data’s problem, which needs further analysis. And there are significant variations between the stage of growth and the amount of VC financing by verifying hypothesis 3. Hypothesis 4 is verified partly, which means the number of startup’s granted patents, the growth stages of financing firms, investors’ funding amount, and VC series will have different characteristics among some, instead of between any two, different industries. So how the feature of different industries make influence on VC should also be studied.

We will make more analysis in the future research. First, in most present research, scholars usually use the number of different types of patents, even don't distinguish the type, to evaluate a firm’s technology innovation. It might be unilateral cause the quality of patent varies. We plan to add text-mining and more indicators of patents to the following research. Second, we will collect more VC and patent data of unlisted companies, who is in more need of VC and will get much more useful direction from this kind of research. Third, patent is only one pattern of manifestations of technology innovation, some other manifestations, such as software, are also innovation of a firm’s technology. So, it is also worthy to consider other patterns of technology innovation.

Acknowledgments
This work was supported by Ministry of Education of the People's Republic of China (No.13YJC630042) and National Natural Science Foundation of China (No.71503020).

References


Tseng, C. Y. (2014). Technological innovation capability, knowledge sourcing and collaborative innovation in Gulf cooperation council countries. *Innovation*, 16(2), 212-223.


The choice of examiner citations for refusals: Evidence from the trilateral offices

Tetsuo Wada

Gakushuin University, Tokyo (Japan)

Abstract
As a comparative study between the EPO, the USPTO, and the JPO, this paper reports on (1) the discrepancies of X/Y patent citations (blocking patents), (2) overall trend of the discrepancies, measured by patent classification vector, and (3) how the discrepancies of X/Y patent citations relate to the characteristics of applications and longitudinal aspect of office actions. Blocking patents both in EP and US commonly show that the first action lag and the breadth of patent class of patent application are positively correlated with divergent refusal reasons, suggesting that costly examination leads to diversified X/Y patent citations for grant/refusal decisions.

Conference Topic
Patent analysis, Analysis on examiner citations, X/Y citations

Background
Patent citations have been widely utilized for empirical studies of patent systems, particularly for such issues as economic value and knowledge flows. Yet, there are numerous kinds of patent citations. Whereas the differences between examiner citations and inventor citations have received better understanding as a result of several empirical studies (e.g., Alacer and Gittleman, 2006; Hegde and Sampat, 2009; Cotropia et al., 2013), there exist few analyses concerning examiner citations specifically for issuing refusals.

Examiner citations for refusals, which are assigned special categories of “X” and “Y” at the European Patent Office (EPO) and at other offices such as the Japan Patent Office (JPO), have important meaning, in that examiners must indicate specific reasons to refuse patent claims, for example, due to lack of novelty. Applicants seek to know which prior art will be used by examiners for refusals, since initial claims are likely to pass through prosecutions to patent grants if examiners do not assign X/Y citations (i.e., if only “A” citations, or simply “relevant” prior art, are assigned). Thus, characteristics of X/Y citations has practical significance for patent prosecution strategies of applicants.

Furthermore, X/Y citations carry crucial piece of information in order to understand the institutional differences between countries. Statutory requirements of novelty and inventive step for granting patents are very similar among advanced economies. Therefore, the same invention described in the same claims and specification should theoretically invoke very close grant/refusal responses from patent offices. However, it is known that the decisions of patent grants and refusals by the trilateral patent offices (the EPO, the JPO, and the US Patent and Trademark Office, or the USPTO) often differ over the same content of triadic patent applications, as was studied by the Melbourne school (Jensen et al. 2005; Webster et al. 2007; Webster et al. 2014). X/Y citations have a clue in evaluating the mechanism by which examiners at the trilateral offices close the same case with different decisions. Yet, direct comparisons between citations for refusals have been challenging, since the US patent system does not provide citation category (such as “X” and “Y”) information. In other words, a novel data set on “blocking patents” in the US must be developed first.
Purpose
This present study takes advantage of a novel large-scale data set of “blocking patents,” obtained from refusal documents available as file wrappers on the Public PAIR database of the USPTO, to compare patent citations employed by examiners specifically as reasons for refusals. In other words, by way of approximating citation categories of X/Y for the USPTO, we are now able to measure technological divergence of individual refusal reasons used by the trilateral patent offices through family-to-family citations.

This proposal set three research questions: (1) the extent of technological discrepancies of patent citations for refusals (blocking patents) between the trilateral offices of the EPO, the USPTO, and the JPO, (2) whether we can find overall trend of the discrepancies, measured by patent classification vector, and (3) how the discrepancies of X/Y patent citations relate to the characteristics of patent applications and the longitudinal aspect of office actions, i.e., the dates on which refusals were presented to applicants.

Data source
We are in the process of developing a large-scale US blocking patent database by extracting patent numbers from the main text of “CTNF” (non-final rejections) and “CTFR” (final rejections) documents available on the Public PAIR database of the USPTO. These document types are different from “PTO-892” (Notice of references cited by examiners) from which US examiner citations are usually extracted, in that examiners state in the main text of CTNF/CTFR which prior arts they actually rely on to reject individual claims. Optical character recognitions and natural language processing are conducted by Guan-Cheng Li, who has also contributed to numerous patent data projects at the University of California, Berkeley. Attorneys at the AIPPI Japan (International Association for the Protection of Intellectual Property, Japan) have also been contributing for this blocking patent database through manual verification. To the best knowledge of the author, there is only one similar, yet separate, large-scale database project on US blocking patents, which is run by Jeffrey Kuhn, based on PAIR documents hosted by Google (Thompson & Kuhn, 2017). A preceding study on blocking patents used only smaller-scale data (Cotropia et al., 2013). Combined with the US blocking patent database, the EPO PATSTAT database (Spring 2016) has been used along with the OECD Triadic Patent database.

The domain of statistical analyses is the set of triadic applications through PCT and non-PCT (initially, 409,653 DOCDB family citation pairs) where their first priority year being between 2003 and 2010 in a patent family. Citing side is a DOCDB family (initially, 43,339 triadic families) where only single DOCDB family ID is observed and where blocking patents are added by all of the trilateral offices. Cited side is also a DOCDB family (initially, 367,054 blocking patent families).

All citation data are patent citations, because of the availability of DOCDB family-to-family citations. It is a weakness, though most observed examiner blocking citations are patents only. We combine citation date information from the EPO DOCDB back file and the USPTO Bulk PAIR data to relate the technological difference with the refusal timing within international patent prosecution processes of an international patent family.
Findings

Simple Discrepancies at the DOCDB family-to-family citation level
First, as Table 1 and Figure 1 show, the trilateral offices largely rely on their own reasons of prior patents (X/Y citations, or blocking patents) for refusals. Based on 409,696 DOCDB family citations found from 43,343 citing families which have single-DOCDB family IDs and priority year 2003-2010, we find surprisingly small share of overlapping refusal reasons, though after the cited patents are consolidated by DOCDB international patent families.

Table 1. Dissimilarity between X/Y citations (“blocking patents,” or examiner citations specified on CTNF/CTFR refusals at the USPTO) at the DOCDB family-to-family citation level.

<table>
<thead>
<tr>
<th>EP_XY</th>
<th>JP_XY</th>
<th>US_XY_EQV</th>
<th>Counts of family-level citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>128,062 31.3%</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>148,459 36.2%</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>11,078  2.7%</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>75,748  18.5%</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>16,541  4.0%</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>19,844  4.8%</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9,964   2.4%</td>
</tr>
</tbody>
</table>

Figure 1. Small overlaps between X/Y citations (blocking patents)

One might wonder that the small overlap is attributable to the prosecutions conducted on non-PCT applications, because non-PCT applications do not carry information from ISRs, which could provide common information of prior art for the designated offices (DOs) conducting national prosecutions. However, as Table 2 shows, the distribution of the non-PCT sample is not much different from the combined sample of PCT and non-PCT. ISRs prepared at the Receiving Offices (ROs) for PCT applications may help searching activities at the DOs, but they apparently do not affect the choice of X/Y citations during national prosecution processes. Because the overlaps of X/Y citations at the trilateral offices are small, we then utilize technological distance measurement to evaluate the substantial differences between them.
Table 2. Dissimilarity between X/Y citations at the DOCDB family-to-family citation level, PCT only

<table>
<thead>
<tr>
<th>EP_XY</th>
<th>JP_XY</th>
<th>US_XY_EQV</th>
<th>Counts of family-level citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>67,877</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>76,192</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>7,179</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>36,537</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>7,732</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>11,270</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5,447</td>
</tr>
</tbody>
</table>

Methodology for comparisons in technological discrepancies

Following Jaffe (1986), we use the cosine distance measurement (similarity index) between bundles of citations using the International Patent Classification (IPC), i.e.,

\[ \text{COS}(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|} \]

where \(\vec{a}\) is a blocking patent family citation vector. An element of this vector is a count variable of all main class assignments of the X/Y citation, i.e., cited patent(s) of the cited DOCDB family, recorded on PATSTAT tis209 (application IPC). No duplicate counts occurred when several IPC subclasses, main groups or subgroups in a main class are assigned, whereas more than one cited family member may result in more than one element in a main class.

In contrast, \(\vec{b}\) is a global citation vector, where a bundle of all citations in the family are added and counted. All DOCDB family members of the bundle are included to count IPC main class, so an IPC main class as an element of the vector may increase when there are many patents cited by the originating application in the bundle, and also when a cited patent has more patent family members. In sum, this measurement indicates how similar a blocking patent family of an office is compared to globally cited (blocking and non-blocking) patents.

If the cosine distance is one, the technological vector of a particular blocking patent is identical with the technological vector of the bundle of all citations added to the same application worldwide. Figure 2 shows an example.

Figure 2. Blocking citation vector \(\vec{a}\) and global citation vector \(\vec{b}\) for a US blocking patent.
**Overall trend**

Utilizing the technology distance associated with a blocking patent, we describe the overall trend of the discrepancies. As is depicted by Figure 3 where vertical axis indicate the cosine similarity index, all blocking patents used by the trilateral offices have a slightly upward trend in recent years. It suggests that the improvement of prior art search technology leads to less heterogeneous blocking patents. Though the USPTO has relatively dissimilar vectors on the average compared to those by the EPO or by the JPO, the overall trend seems to be the same.

![Figure 3. Cosine distances of refusal reasons by the trilateral offices from all citation vector, by priority year.](image)

**Complexity, longitudinal distribution, and technological distances of blocking patents**

Unlike citations made on academic papers, examiner citations for refusals can be issued several times during a prosecution in response, for example, to amendments. In other words, examiner citations are produced in a dynamic process of bargaining between an applicant and an examiner. Therefore, longitudinal aspect of blocking citations should be an important factor as one of the determinant for technological distance, especially in the US where repetitive refusals are frequently observed. The US blocking patents are sometimes added by examiners up to ten times over more than ten year period from application in the same sample above.

Also, the discrepancies of X/Y patent citations may relate to the technological complexity of patent applications. This is because examiners at different offices might focus on diverse aspects of a patent application if the application is technologically complex and has a broad scope.

In order to explore the relationship between technological discrepancies in blocking citations and their longitudinal aspects further, linear regression analyses are implemented, taking the cosine similarity measurement as the dependent variable. Several explanatory variables are employed. The first group represents longitudinal aspects. The time lag in years between an entry of a triadic application into the US and the first rejection action by the USPTO is given by `us_firstaction_lag`. Similarly, the time lag in years between an entry of a triadic application into the EPO and the EP search report date is given by `ep_action_lag`. If an
application is technologically complex, this lag variables are expected to increase, and we expect these to be related negatively with the divergence of refusal reasons in the US and at the EPO, respectively, because examiners will focus on different features of the same application and will cite different prior arts. Following the first action, one or more additional office actions might take place at the USPTO. The lag in years between the first action and the following office action is given by \textit{repeated\_action\_lag}. Since repeated actions often occur within limited ranges of claim issues, and since the difference between offices decreases as examiners in different locations have similar information, we expect this to be positively related with the divergence of refusal reasons. Unfortunately, office action timing is not available at the EPO by PATSTAT and by DOCDB. Therefore, we are not able to define a similar variable for the EPO. Instead, we define \textit{ep\_us\_lag\_diff\_abs}, which represents the absolute time difference in years between the USPTO office actions and the EPO search reports. When this value is large, examiners at the USPTO and the EPO tend to face different information concerning the application. Therefore, we expect that there is negative relationship between this variable and the similarity index of blocking patents.

The technological complexity of a triadic patent application is represented by the number of 35 WIPO technological fields assigned to the application, or \textit{techn\_field\_nr\_counts}. We expect this variable to have negative relationship with the similarity of a blocking patent depicted by examiners. In addition to the explanatory variables, we employ control variables of other characteristics. Specifically, \textit{isr\_cited\_dummy} is a dummy variable indicating the blocking patent was shown in the preceding international search reports. Whether a blocking patent is used at other offices is coded as binary variables of \textit{ep\_xy} and \textit{jp\_xy}, which mean the blocking patent is also X/Y citation at the EPO and at the JPO, respectively. Patent grants ex post are controlled by three dummy variables of \textit{epo\_granted}, \textit{us\_granted}, and \textit{jpo\_granted}. In addition, priority years and technological fields are controlled.

As linear estimation results in Table 3 indicate, both the US and the EP blocking patents commonly show that the first action lag, the time differences between actions at the USPTO / the EPO, and the breadth of patent class of patent application (\textit{techn\_field\_nr\_counts}) are negatively correlated with divergent refusal reasons. It is consistent with an interpretation that costly examination leads to diversified reasoning for grant/refusal decisions. However, office action lag after the second action (\textit{repeated\_action\_lag}, where data is available only for the US at this time) shows the opposite, suggesting that repetitive refusals tend to be focused on a limited range of issues in terms of prior art. The results suggest that blocking patents at the EPO and the USPTO become more divergent as examination is more costly, which requires longer time for prosecution.

Table 3. Linear regression, dependent variable: cosine distance between blocking patent vector (US/EP) and global citation vector. Unit of analysis: DOCDB family citation pairs

<table>
<thead>
<tr>
<th>Dependent Var.</th>
<th>Cosine distance between US blocking patent vector and global citation vector</th>
<th>Cosine distance between EP X/Y citation vector and global citation vector</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
</tr>
<tr>
<td>\textit{us_firstaction_lag}</td>
<td>-0.0036086*** (0.0010477)</td>
<td>-0.0026007** (0.0010607)</td>
</tr>
<tr>
<td>\textit{repeated_action_lag}</td>
<td>0.0076326*** (0.0011132)</td>
<td>0.0010688 (0.00168)</td>
</tr>
</tbody>
</table>

955
Robust standard errors in the parentheses. ****<0.001, ***<0.01, **<0.05
35 WIPO technology field dummies are included and controlled for citing families.

**Conclusion**

As a conclusion for the first question of the extent of technological discrepancies of blocking patents, we have found that the decision reasons used by examiners are very dissimilar between trilateral offices, if we define similarity by coincidences between DOCDB patent families of blocking patents. Secondly, concerning the second research question, overall trends in the similarity between blocking patents suggest that blocking patents become less divergent at the trilateral offices recently, possibly due to the advancement in search technology. Converging rejection reasons may lead to less divergent grant decisions between the trilateral offices. Thirdly, as a finding for the third question, we have found evidence which is consistent with an understanding that blocking patents at the EPO and the USPTO become more divergent as examination burden is heavier. This is consistent with Wada (2016) in that examiners are not free from the limitation in searching capability for prior arts, i.e., examiners are constrained in their information space.
Our understanding of the driving factors for the dissimilarity is still incomplete. For example, there are many other methods to measure the technological discrepancies (Wang, Thijs, & Glänzel, 2015), whereas this paper has utilized Jaffe’s cosine distance measurement only. Also, due to data limitation of office action timing, we have not been able to employ action lag data for the JPO. Although we have not obtained enough information about the details of patent applications and bias made by examiners, this paper has added knowledge to the understandings of limitations and biases of patent prosecution processes. We now have initial clues to analyse how examiner citations are produced and biased by way of measuring divergent blocking patent patterns.

Acknowledgments

The author acknowledges the financial support of the Japan Science and Technology Agency under RISTEX Project “Development of Benchmarks of the Quality of Prior Art Search in International Patent Prosecution Processes,” and the Research Institute of Economy, Trade and Industry (RIETI). The author would like to thank two anonymous referees for the ISSI Wohan conference, Professor Sadao Nagaoka, Setsuko Asami, Yoshimi Okada, Hisao Shiomi, Stuart Graham, participants of the 2016 Annual Conference of the Asia-Pacific Innovation Network for helpful comments.

References

Interactive overlay maps and technological field evolutions for US patent (USPTO) data

Huang Ying¹ Zhu Donghua² Qiao Yali³ Pang Jinhui ⁴ Wang Xuefeng ⁵,*

¹huangying_work@126.com; ²zhudh111@163.com; ³qiaoyali_cnu@163.com; ⁴pangjinhui@bit.edu.cn; ⁵wxf5122@bit.edu.cn (corresponding author)

Beijing Institute of technology, Beijing (China)

Abstract

Detecting the technical information included in patent documents offers a window of opportunity for identifying current technology structures and tracing technological developments. In this paper, we attempt to build a basemap for the international patent classifications (IPCs) of USPTO citation data, with the aim that such a map might be used as a tool to benchmark and capture the changes in organizational innovation activities over time, or the patterns in technological changes for future investigations in other sectors. Additionally, we subdivided patent classifications into detailed technological fields, as opposed to the usual eight broad technical domains, then traced the direction of the information flow over time. Promising quantitative measures, based on the technological proximity of the specialization and integration of the classes, were introduced to further provide insights into the notable differences in technological change and knowledge fusion over the past 30 years. Our aim in constructing this interactive overlay, and tracing the evolution of these technological fields, is to understand the patent portfolios of countries, organizations, and new and emerging technologies in the context of the overall technological landscape.

Conference Topic

Mapping and visualization, Patent analysis

Introduction

Patents, as important carriers of technological knowledge, can be used to analyse patterns of invention by location, technology class, or organization. As compilations of large numbers of records for “landscaping”, patents provide both a macro perspective and the opportunity for in-depth, micro-analysis of a small number of patent documents (Trippe 2003). Examining the technical information included in patent documents offers a window of opportunity to identify current technology structures and trace technological developments. With keen observation, one may be able to spot the drift and development trends in many commercial sectors.

The visualization techniques that transform abstract and intangible data sets into something visible and concrete provide us the perceptual and cognitive ability to grasp the complexity and evolution of scientific and technological activity patterns more easily and more vividly than traditional scientific studies (Chen and Kuljis 2003). Among these techniques, the science overlay maps (Leydesdorff and Rafols 2009; Rafols et al. 2010) that identify scientific domains, analyze change in scientific frontiers, and inspire exploration of interdisciplinary scientific progress open up new avenues for understanding patent landscapes.

Technology classes are widely considered to be an appropriate unit of analysis for measuring technological proximity (Ejermo 2005). Technology-oriented applications often draw from patents in different hierarchical categories, and this subsequently leads to further diversity in patents that cite patents in these categories. There are three major classification systems: the International Patent Classification (IPC) system developed by the World Intellectual Property Organization (WIPO) in Geneva; the United States Patent Classification (USPC) system used by the US Patent and Trade Office (USPTO), and the European classification system (ECLA) refined by the European Patent Office (EPO) in partnership with WIPO and the Organization for Economic Cooperation and Development (OECD). The Cooperative Patent Classification (CPC) system, a new classification that covers all EPO and US classified documents, is
thought to be the most granular and precise classification system of the English versions (Montecchi et al. 2013). However, CPC and IPC are identical in the first four characters, while IPC is still the most widely used by most countries in the world.

Among national patents, the USPTO’s patents are considered the most accessible and yet also the most valuable. The availability of these data and the length of their history means they have been applied to many studies. Moreover, non-US firms have persistent interests in protecting their most significant innovations in the US because it is one of the most highly competitive markets for technologically advanced products (Criscuolo 2006).

In this paper, we attempt to build a basemap of IPC for USPTO citation data, with the aim that such a map might be used as a tool to benchmark and capture the changes in organizational innovation activities over time, or the patterns in technological change for future explorations in other sectors. Additionally, we subdivided patent classifications into detailed technological fields rather than the eight broad technical domains used in IPC, and then traced the direction of technological information flow over time. Promising quantitative measures based on the technological proximity of the specialization (S) and integration (I) of the classes were introduced to provide further insights into the notable changes in technologies and knowledge fusion over the past 30 years.

Data

To build this type of interactive patent overlay map, one of the toughest issues we faced was how to obtain high-quality patent data with equivalent citation patterns in appropriate quantities. Unlike the Web of Science categories applied in many science overlap maps, significant variations in the number of patents classified within each IPC classification level increase the challenges brought by the intensive quantity of patents in certain categories.

Our original patent data was collected from the National Bureau of Economic Research (NBER), a research group that has been working to make US patents widely accessible for research over the past decades (Hall et al. 2001). We selected the patents granted between 1976 and 2005 for this study, which resulted in 3,035,457 utility patents. Technical patents, design patents and genetic sequences were excluded. For the best visualization, and so as not to lose necessary detail or resolution of the technology landscape, we chose 4-digit classes that met the degree of detail in subject matter definition to represent network vertices. Ultimately, a dataset of 3,015,846 patents spanning 630 4-digit IPCs were selected as the classifications to benchmark in keeping with Leydesdorff et al. (2014) study.

The number of application patents and granted patents by publication year is presented in Figure 1. It can be seen that both the applications and granted patents show notable growth over the past several decades and that more and more patent assignees have been treating patent applications as an effective approach to protecting their intellectual property. The fluctuating number of patents granted and the relatively smoother trend for application numbers is another notable characteristic. The tremendous growth in patent applications has led to a noticeable delay in the examination process. Patent applications were granted relatively quickly before the 1990s, and, after 2001, a far smaller proportion of applications have been granted.
Even though 630 4-digit IPCs are more than adequate to compare the number of patents belonging to each subclass, classifications based on too much detail are meaningless in statistical analysis, and of course, classifications that are too broad would cover too many technologies to appropriately distinguish. One proposed solution for reclassifying the subclasses of patent categories is to fold these 4-digit categories up into higher levels of the appropriate number. However, in 1992, Fraunhofer ISI and the Observatoire des Sciences et des Technologies, in cooperation with the French patent office (INPI), developed a more systematic technology classification, called the ISI-OST-INPI classification, based on the IPC codes (Schmoch 2008). This classification system has been amended several times to keep pace with the revolution of IPCs, and new codes have been added. The latest edition from the WIPO Statistics Database includes 35 technological fields that cover and balance all possible IPCs (an Excel spreadsheet is available at: www.wipo.int/ipstats/en/statistics/patents). However, it is worth noting that these codes are exclusive to IPCs and quite distinct from each other.

The ISI-OST-INPI system provides a feasible manner for clustering our 630 4-digit IPCs into interactive overlap maps, but it has two main limitations. First, most IPCs are clustered at the subclass level as 4-digit IPCs, but some are clustered at a group level which sits even lower at the fourth hierarchical level of the system. Next, the number of IPCs in each different field are uneven, and this does not benefit the structure of the map. We took these 35 fields as benchmark technological fields and merged them into 29 optimized fields according to the following three rules: (1) All IPCs at a group level were upgraded to their subclass level. If different group levels belonged to different subclasses, we chose the one that contained the most groups. (2) Fields with less than 5 IPCs were merged into their most relevant field according to the strength of its citations. For example, A61K and A61P previously belonged to Pharmaceuticals but were merged into Medical Technology; B81B, B81C, B82B and B82Y were grouped into Chemical Engineering from Micro-structural Engineering and Nano-technology; G06Q (IT Methods for Management) was merged into Computer Technology; H01L (Semiconductors) was merged into Electrical Machinery, Apparatus, and Energy; H04L and H04W clustered under Digital Communication were merged into Telecommunications. The number of patents in each corresponding technological sector and field are provided in Table 1.
Table 1. The number of patents in each technological sector and their corresponding fields

<table>
<thead>
<tr>
<th>Sector</th>
<th>Fields</th>
<th>Number of Patents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemistry (939,644)</td>
<td>Biotechnology</td>
<td>90,361</td>
</tr>
<tr>
<td></td>
<td>Chemical Engineering</td>
<td>127,101</td>
</tr>
<tr>
<td></td>
<td>Environmental Technology</td>
<td>25,749</td>
</tr>
<tr>
<td></td>
<td>Food Chemistry</td>
<td>28,121</td>
</tr>
<tr>
<td></td>
<td>Macromolecular Chemistry</td>
<td>105,228</td>
</tr>
<tr>
<td></td>
<td>Materials Chemistry</td>
<td>92,739</td>
</tr>
<tr>
<td></td>
<td>Materials, Metallurgy</td>
<td>72,146</td>
</tr>
<tr>
<td></td>
<td>Medical Technology (Pharmaceuticals)</td>
<td>114,558</td>
</tr>
<tr>
<td></td>
<td>Organic Fine Chemistry</td>
<td>160,151</td>
</tr>
<tr>
<td></td>
<td>Surface Technology</td>
<td>123,490</td>
</tr>
<tr>
<td>Electrical Engineering (895,879)</td>
<td>Audio-visual Technology</td>
<td>149,191</td>
</tr>
<tr>
<td></td>
<td>Basic Communication</td>
<td>66,254</td>
</tr>
<tr>
<td></td>
<td>Computer Technology</td>
<td>212,408</td>
</tr>
<tr>
<td></td>
<td>Electrical Machinery</td>
<td>335,120</td>
</tr>
<tr>
<td></td>
<td>Telecommunications</td>
<td>132,906</td>
</tr>
<tr>
<td>Instruments (567,314)</td>
<td>Control</td>
<td>53,254</td>
</tr>
<tr>
<td></td>
<td>Measurement</td>
<td>200,577</td>
</tr>
<tr>
<td></td>
<td>Medical Technology</td>
<td>142,998</td>
</tr>
<tr>
<td></td>
<td>Optics</td>
<td>170,485</td>
</tr>
<tr>
<td>Mechanical Engineering (901,846)</td>
<td>Engines, Pumps, and Turbines</td>
<td>104,089</td>
</tr>
<tr>
<td></td>
<td>Handling</td>
<td>138,258</td>
</tr>
<tr>
<td></td>
<td>Machine Tools</td>
<td>127,519</td>
</tr>
<tr>
<td></td>
<td>Mechanical Elements</td>
<td>132,362</td>
</tr>
<tr>
<td></td>
<td>Other special machines</td>
<td>123,316</td>
</tr>
<tr>
<td></td>
<td>Textile &amp; Paper Machines</td>
<td>89,340</td>
</tr>
<tr>
<td></td>
<td>Thermal Processes &amp; Apparatus</td>
<td>52,498</td>
</tr>
<tr>
<td></td>
<td>Transport</td>
<td>134,464</td>
</tr>
<tr>
<td>Other fields (292,855)</td>
<td>Civil Engineering</td>
<td>106,657</td>
</tr>
<tr>
<td></td>
<td>Furniture, Games</td>
<td>105,088</td>
</tr>
<tr>
<td></td>
<td>Other consumer goods</td>
<td>81,110</td>
</tr>
</tbody>
</table>

Methods

Constructing the basemap for the patent overlay maps

Patents that reference other patents in related technological fields have been suggested as a way of offering access opportunities to advance an area, whereas patent documents that contain references across diverse categories may offer the potential for radical innovation (Olsson 2005). In general, the IPC class of every patent might range from one to more than twenty, in exceptional cases. As the first step in constructing a basemap for patent overlap maps in our aggregated set, we used the cross-citation patterns of 630 4-digit IPCs of granted patents during the period 1976-2005. Given most patents relate to more than one field, co-classifications have been used more often than multi-classifications to measure the “technological distance” between classes (Breschi et al. 2003). However, citation networks are less sensitive to misclassifications than co-classification networks because they offer an additional degree of freedom compared to hierarchies. This means that the actual number of papers per overlay map depends on the breadth of the disciplinary category distribution (Rafols et al. 2010). As a result, we chose cross-citation rather than co-classification to build a
citation network, and reveal the structural patterns of technological information flow among the 630 4-digit IPCs.

Constructing this type of patent network requires a proper measure of the distance between different classes of patents in the patent classification system (Yan and Luo 2017). Therefore, the second data processing step involved normalizing the similarity matrix for the cross-citation IPC categories into an aggregated citation matrix. Initially, a correlation between the vectors representing the distribution of a firm’s patents in a set of technology fields was used as the index for technology mapping (Jaffe 1986). However, for co-citation patterns, Salton’s cosine measure (Salton and McGill 1986) is suggested instead of Pearson’s correlation coefficient, particularly if one aims to visualize the network’s structure, as in the case of social network analysis or multidimensional scaling (MDS) (Ahlgren et al. 2003). A cosine-normalized asymmetrical occurrence matrix provides the best visualization of vector space and is more consistent with consensus science mapping. The cosine-normalized approach uses conventional cosine similarity normalized by the square root of the squared sum. The cosine of the angle between the two vectors represents the citation distributions between two technology classes among all patent classes (Leydesdorff 2007). The formula can be presented as

\[
\text{Cosine} (i,j) = \frac{\sum_{k=1}^{n} C_{ik} C_{jk}}{\sqrt{\sum_{k=1}^{n} C_{ik}^2 \sum_{k=1}^{n} C_{jk}^2}}
\]

where \( C_{ij} \) indicates the number of citations referred from patents in technology class \( i \) to the patents in technology class \( j \) and \( k \) belongs to all the technology classes. The cosine value falls between \([0,1]\) and indicates the similarity of the knowledge bases of the two fields.

The remaining question that needed to be taken into consideration is which overlap toolkit to use. There are three main toolkits that offer a wealth of possibilities for the visualization of a comprehensive map and enable us to visualize portfolios as overlays: Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/), Gephi (https://gephi.org/) and VOSViewer (http://www.vosviewer.com/). Pajek and Gephi contain a suite of tools for network analysis and visualization, such as various decompositions, layouts, and visualization options. But, the advantage of VOSViewer is that further normalization within this toolkit does not disturb the maps on the basis of cosine-normalized matrices (Leydesdorff and Rafols 2012), and this feature drove us to choose VOSViewer as the benchmarking visualization platform. Technological fields are positioned in the map so that similar fields are situated nearby, and dissimilar components are situated at a distance.

**Mapping technological information flow among different fields**

The duality of citations and references is a central feature of citation processes. In terms of the direction of technological information flow, referencing is regarded as emission and the citations as reception. In order to map technological information flow among 29 different technological fields, rather than 630 4-digit IPCs and to distinguish the extent of the information flow, we introduced a new variable – citation link strength (CLS) (Zhang et al. 2009). The formula for the CLS between the technological fields \( i \) to \( j \) can be presented as

\[
\text{CLS}_{ij} = \frac{a_{ij}}{\sqrt{T Ci} \cdot TR_j}
\]

Here, \( i \) and \( j \) denote technological fields, \( TCi \) represents the total number of citations of the field \( i \), \( TR_j \) denotes the total number of references of the field \( j \), and \( a_{ij} \) denotes the number of citations the field \( i \) receives from field \( j \).
This indicator measures the strength of the citation links between two technological fields in an asymmetric matrix, which is directional as citations from field \(i\) to field \(j\) differ from citations from \(j\) to \(i\). Through this indicator, we can compare the emissions (the referencing) and the receipts (the citations) of an individual technological field, and can also detect some extremely asymmetric links between two technological fields. The footprints of the technological information flow among different fields also provide different dimensions for understanding the complex dynamics in innovation by providing different projections of those dynamics.

**Measuring specialization and integration in technology**

A large number of technologies represent combinations of different research fields and patented technologies, so it is important to examine the extent of diversity in the underlying research fields and the patented technologies to understand the inherent composition of these fields. We introduced two indexes, an integration index and specialization index, to measure the technological change in patents to more comprehensively navigate the evolution of a technology.

Traditionally, the interdisciplinary metrics of integration and specialization are based on ISI subject categories and measure the interdisciplinarity in an aggregated dataset of a country, institution or individual researcher (Porter et al. 2008; Porter and Rafols 2009; Porter et al. 2007). In this paper, we introduce a specialization score and an integration score, comparable to the Rao-Stirling’s diversity index (Rao 1982; Stirling 2007), to measure technological diversity in patenting. Likewise, the integration score can be used to examine the spread of references in a given technological fields to gauge the degree of integration across “bodies of specialized technology” reflected by the span of IPCs. The formula can be presented as

\[
\text{Integration} = 1 - \frac{\sum C_i C_j S_{ij}}{\sum C_i C_j}
\]

Here, \(C_i\) indicates the number of citations received by technology class \(i\), \(C_j\) represents the number of citations received by technology class \(j\), and \(S_{ij}\) is the technological similarity between categories \(i\) and \(j\). The integration score, ranging from 0 to 1, increases as a technological field cites more disciplines. Scores closer to 1 indicate more integration.

Specialization is operationalized differently from integration in important ways, reflecting the breadth of technology field distributions. The formula can be presented as

\[
\text{Specialization} = \frac{\sum F_i F_j S_{ij}}{\sum F_i F_j}
\]

Here, \(F_i\) indicates the frequency of technology class \(i\), \(F_j\) represents the frequency technology class \(j\). Specialization can reflect the breadth of technological fields over a certain time span.

**Results**

**Interactive overlay maps based on 630 4-digit IPCs**

The overlap patent basemap in Figure 2 is based on cross-citation relationships among the 630 4-digit IPCs of USPTO patents from 1976 to 2005. The 29 technological fields listed in Table 1 form the basis for color-coding these 630 categories. Labels and colours have been used to produce a reasonably clear map and facilitate its examination.
One good application for patent overlay maps is their use as benchmarking tools for companies or specific technological fields. To illustrate the application of patent overlay maps, a data set was created for patents granted on graphene, using data from the USPTO for the publication years 2000–2015. As shown in Figure 3, the subfields relating to graphene are mainly focused on Electrical Machinery, Surface Technology, Chemical Engineering, and Material Metallurgy. Hence, applying patent overlays to the analysis of technological subfields can help provide a better understanding of technologies involved in the development of a subfield.

Figure 2. The basemap of 630 4-digit IPCs

Figure 3. The patent overlay maps for graphene from the USPTO during the period 2000-2005
Note: The dataset is retrieved using the term “Graphene” in the fields of Title and Abstract in USPTO granted patents, and result in 1003 patents.

Technological information flows for 29 main technological fields

Compared to individual IPCs, analysis of the direction of information flows among different technological fields provides views at a macro-level. On the basis of the 29 technological fields defined above, we detected some distinct asymmetric links between different fields. To trace the evolution of technological flows among these different fields, we divided these patents into four time windows according to their publication year: 1976-1985, 1986-1995, 1996-2005, and 1976-2005. The thickness of the lines represents the intensity of the citations between two linked fields, and the size of the nodes is proportional to the number of patents in a given field. A pair of fields are considered as having strong asymmetric links if $\text{CLS}_{ij} - \text{CLS}_{ji} > 0.02$. The technological flows among these 29 fields can be viewed in Figure 4, generated from Gephi.

![Figure 4. The technological flow among 29 fields during four different periods](image)

During the late ‘70s and early ‘80s, patents in Electrical Machinery were somewhat dominant compared to others. The categories are relatively concentrated given they are typically related. In this era, four technological flows are remarkable: from Biotechnology to Organic Fine Chemistry; from Telecommunications to Computer Technology; from Organic Fine Chemistry to Medical Technology; and from Biotechnology to Food Chemistry.
From 1986 to 1995, due to advancements in technology and improved awareness of protecting intellectual property, most fields show an increased number of patents. Patents in Electrical Machinery and Medical Technology are still dominant over others. However, unlike before, there are more links among the fields, and some have links to more than three disciplines. Electrical Machinery shows expansive connections to other fields and is absorbing a great deal of knowledge from external fields, especially Materials Metallurgy. Computer technology was undergoing rapid development during this period, and Computer Technology also shows more and more connections with other fields, heavily relying on Audio-visual Technology. Moreover, Organic Fine Chemistry and Chemical Technology received many more citations than they emitted, while Chemical Engineering emitted many more than it received.

Between 1996 and 2005, the majority of patents were assigned to Computer Technology, Medical Technology, and Electrical Machinery. Electrical Machinery has far fewer links to other fields than before, implying a stronger independence and specialization. Computer Technology has become strongly related to Telecommunications. Most obviously, Biotechnology has become more aligned to Organic Fine Chemistry, Food Chemistry, and Measurement, revealing great advancements in biotechnology. Other knowledge flows between Organic Fine Chemistry and Medical Technology, and between Medical Technology and Materials Chemistry can also be observed.

In summary, throughout this entire period, most emerging patents were distributed in the Electrical Machinery, Medical Technology, and Computer Technology fields. Computer Technology and Electrical machinery relied heavily on other fields, which changed over time. Based on a correlation of the research, Chemical Engineering, Macromolecular Chemistry, Organic Fine Chemistry, Medical Technology relate to each other. A comparison of the citations between the different periods reveals several insights. First, we can trace the underlying development of emerging technology in some fields. For example, Computer Technology began to develop in the period 1986-1995, while Biotechnology has developed rapidly over the past decade. Second, we can identify changes in the basic fields that a technology relies upon. Computer Technology absorbed much knowledge from Control in 1976-1985, then relied more on Audio-visual Technology in 1986-1995, and Telecommunications in 1996-2005. In addition, Electrical Machinery transformed from high dependence to specialization over the period under study.

Technology specialization and technology integration among different sectors

We collected the IPCs for each record and counted them by publication year from 1976 to 2005. These data were represented as a vector denoting the number of IPCs assigned per year to five different fields – Chemistry, Electrical Engineering, Instruments, Mechanical Engineering and other fields. Using these vectors, we calculate the integration and specialization indexes at an aggregate level by patent publication year. Figure 5 summarizes the annual trends in technology specialization and the technology integration for all fields.

We derived two observations from Figure 5. First, the overall patent specialization index was stable at first, and increased slightly in the following years but always remained at a low level. This indicates that the IPCs involved in these patents are technologically distant. Second, the technology integration for all fields fluctuated before 1987, then exhibited a slightly decreasing tendency, which reveals a diminishing capacity to bring together technology from disparate fields.
To better distinguish between the technology integration and specialization among different sectors, we illustrated the annual trends of the technology specialization and technology integration in four different technological fields, as shown in Figure 6. This comparison shows several things. First, the annual specialization and integration scores both reveal fluctuating patterns in the Chemistry field. In the Electrical Engineering field, the specialization score increased rapidly, while the integration score declined slightly. Accordingly, Instruments reveals an increasing tendency in specialization and a drop in integration. As for Mechanical Engineering, neither score has changed much over time. Second, Mechanical Engineering shows a higher degree of integration and a lower degree of specialization relative to the other three fields across the given period. The most obvious difference in the index among the four fields is that patents in Electrical Engineering are more likely to specialize in fewer technological application fields.

Figure 5. The overall technology Integration and Specialization during the period of 1976-2005

Figure 6. Technology integration and specialization of four sectors during the period 1976-2005
Conclusion
The use of patent information increases the technical capabilities of obtaining competitive intelligence, and the patent mapping offers an approach to creating a graphical or physical representation of the relevant art pertaining to a particular subject area or novel invention. Interactive patent overlay maps can be an effective approach for understanding the overall landscape of technological areas and grasping the extent and evolution of technological activities over time.

This paper presents the preliminary results of a new patent visualization tool that has the potential to support policy decision making. The approach involves a two-step visualization process. First, we collected patent citation information and constructed a patent citation network that shows the technological distance among patent categories using USPTO granted patents published during the period of 1976-2005. Second, we overlaid the patent activity of specific technological fields over the fixed “backbone” of a patent map. Such patent overlay maps can be applied to benchmark organizations, regions or countries and form a general analysis of technological changes over time.

The strength of citation links among different technological fields in different periods reveal that technological information flows fluctuate, and the focal technological fields are transformed with the development of human society. Detecting the evolution of technological fields is beneficial for grasping the strategic intelligence required by analysts and decision makers.

Two advanced diversity measurements, specialization and integration, yield different implications when interpreting the interdisciplinarity (or diversity) of a given body of knowledge. A patent specialization score measures how much a developing technology is related to a particular technological application, and a patent integration score indicates how many technological fields have merged to navigate the innovation trajectories of a technology (Kwon et al. 2016).

Acknowledgments
We acknowledge support from the General Program of the National Natural Science Foundation of China (Grant No.71673024). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the supporters or the sponsors.

References


Productivity versus citation impact:
A study of persons, not just authors

Lin Zhang¹,²  Gunnar Sivertsen³

¹Zhanglin_1117@126.com
Dept. Management and Economics, North China University of Water Resources and Electric Power, Zhengzhou (China)
Centre for R&D Monitoring (ECOOM) and Dept. MSI, KU Leuven 3000 (Belgium)
³gunnar.sivertsen@nifu.no
Nordic Institute for Studies in Innovation, Research and Education, Oslo (Norway)

Abstract
In two independent studies that were first presented at ISSI 2015 and recently published in PLOS ONE, Larivière & Costas (2016) and Sandström & van den Besselaar (2016) observe similarly that productivity among individual researchers is correlated with citation impact in large datasets from Web of Science (WoS). While the latter study draws the policy implication that productivity should be incentivized, the first study explains their finding by the Mertonian theory of cumulative advantages and maintains that research assessment should be qualitative and focus on research quality. Both studies are based on author name disambiguation and Web of Science data. As acknowledged by the authors, there are several problems with studying individual productivity by using author name disambiguation. In our study of the same general research question, we instead match Web of Science with national datasets that have information about the persons behind the author names. We know their age, gender, position and affiliation, as well as their former career and educational background in the higher education sector. We also have bibliographic information about all their peer reviewed scholarly and scientific publications – not only those covered by Web of Science. Our aim is to get deeper insight into the conditions that influence observable productivity differences among authors in Web of Science, and thereby to establish a firmer basis, both for understanding the correlations found in the former studies, and for expressing the implications for research policy.

Conference Topic
Science policy and research assessment; Studies on the level of individual scientists

Introduction
In two independent studies that were first presented at ISSI 2015 and recently published in PLOS ONE, Larivière & Costas (2016) and Sandström & van den Besselaar (2016) observe similarly that productivity among individual researchers is correlated with citation impact in large datasets from Web of Science (WoS). The policy implications drawn from this observation are different, however. Larivière & Costas (2016) frame their study within the increasing literature that warns against the possible adverse effects of quantitative research assessment and funding methods. They study whether the incentive to publish as much as possible leads to lower citation impact, but find, on the contrary, that higher productivity does not influence citation impact negatively. They explain the result as a confirmation of the Mertonian theory of cumulative advantages in research and nevertheless maintain that, in line with the Leiden Manifesto (Hicks et al., 2015), individuals are best assessed by qualitative peer review. Sandström & van Besselar (2016) also relate their discussion to the Leiden Manifesto and the policy discussion about a possible overproduction and decrease in research policy. Here, they take an opposing position and conclude that increased productivity in the research system is not a perverse effect of output oriented evaluation systems, but a positive development. Interpreting quantity as the number of WoS-articles and citation impact as quality, they found that there is no evidence that quantity and quality are opposed to each other. Instead, their view is that evaluation methods based on peer review of only a few publications “disadvantages the most productive and best researchers”.

In our research in progress, we will study persons, not only authors, to go deeper into the possible explanations for the observation that the two studies have in common, thereby securing a better basis for the interpretation of the phenomenon and for drawing policy implications.
A common feature of the two studies is that they identify individual researchers by using *author name disambiguation* within data from WoS, and that they do not go beyond this database for further information about the researchers. As an example, Lariviè re & Costas (2016) determine the ‘age’ of the researcher by the time the author name first appears in Web of Science, while Sandström & van den Besselaar (2016) determines the affiliation of a researcher by using disambiguated author addresses. Both studies discuss some of the problems with relying on WoS only: Most publications are co-authored between researchers who contribute in different roles and with varying degrees of efforts, and some of them may have shorter careers as authors. Productivity also differs by gender. In some fields, particularly in the social sciences and humanities, Web of Science does not represent the full productivity of researchers.

There may be other problems with the method as well: The type of institution and the position of the researcher may influence the time and resources available for research. Tenure and external funding can also make a difference. Co-authorship practices and productivity are variable within the five large research areas (applied sciences; natural sciences; health sciences; economic & social sciences) used for field normalization of productivity by Sandström & van den Besselaar (2016). The same is true for four large research areas (law, arts, and humanities; medical and life sciences; natural sciences; social and behaviour science) used to control for different publication patterns by Lariviè re & Costas (2016). Sandström & van den Besselaar (2016) are more advanced in controlling for average productivity in a field. They calculate *Field Adjusted Production* (FAP) with the so-called Waring method, but it has been documented to suffer from significant flaws in quite similar data representing all Swedish authors (Vetenskapsrådet, 2010).

We will try to come around as many as possible of such problems by selecting comparable data from WoS and replace the author name ambiguation method by matching the WoS data with two other data sources at a national level where the author names and addresses in the publications can be linked to real persons and institutions. From these other data sets we will also know the age, gender and position of the researchers, as well as their former career and educational background in the higher education sector. In addition, we will have complete bibliographic information about all their peer reviewed scholarly and scientific publications – not only those covered by WoS. This information is possible to gather at the national level in Norway by matching data from Web of Science to other datasets, as explained below and demonstrated in previous publications, e.g. Aksnes et al. (2013) and Zhang et al. (2017).

This is a research in progress paper because the large task of matching the data sets remains to be done. Our final results will be presented at ISSI 2017 if we are given the possibility to do so. However, our preliminary investigations, as demonstrated below, indicate the we will observe productivity and citation impact as correlated with mostly the same pattern and to the same degree as in the two above mentioned studies. Some of the methodological problems in the two former studies will be solved at the same time, thereby laying a firmer basis for empirically based conclusions.

As discussed above, the two former studies arrive at different policy implications. We believe that this may partly be due to the lack of data on the researchers as persons in both studies. With this data, we will get deeper insight into the conditions and factors that influence observable productivity differences among authors in Web of Science. The performances of researchers, as measured within Web of Science with bibliometric indicators after disambiguation of author names, may be interrelated in a complex way with what roles researchers actually take in research, what position they have, what resources are available, and what they achieve in their careers. We expect to be able to provide a better understanding of what the bibliometric measurement of productivity and citation impact might mean at the intermediate level, thereby providing a firmer basis for drawing policy implications from individual level studies.
Norway is mentioned by Larivière & Costas (2016) as a country with an institutional funding system based on bibliometrics, thereby as an example of a country where incentives for increased productivity may lead to lower research quality and other adverse effects. There have been several studies of the effects of the Norwegian system recently (e.g., Aagaard et al., 2015; Bloch & Schneider, 2016; de Rijcke et al., 2016; Sivertsen, 2016). None of these studies, however, have used individual level data sources with the same strategy as in our study. By using Norwegian data, we thereby also hope to enlighten the debate in which Larivière & Costas (2016) and Sandström & van den Besselaar (2016) draw their different policy implications.

Data

Our study depends on the matching of three data sets at the individual article and author/researcher level:

- **National Citation Report for Norway (NCR, 1981-2016)**, a data set provided by Clarivate Analytics with a representation of all articles in Web of Science with minimum one address in Norway, and their accumulated citation counts. This data set has the same field classifications and indicators that are present in the product InCites from Clarivate Analytics, but it is delivered as a database with all the basic data.
- **The Norwegian Science Index (NSI)**, a subset of the Current Research Information System in Norway (Cristin), with complete coverage since 2010 of all peer-reviewed scientific and scholarly publication output, including books, edited volumes, and conference series (Sivertsen & Larsen, 2012).
- **The Norwegian Research Personnel Register**, which has been updated since 1977 and has individual data for all researchers at public research institutions in Norway, including age, gender, educational background, affiliations and positions.

Publications from the five-year period 2010-2014, with citations until the end of 2016, will be selected in all fields from Norway’s four largest universities (Bergen, Oslo, Trondheim, Tromsø). The rest of the higher education sector is not included because researchers at other institutions on the average have less time and resources for research activities.

There are 34,417 research articles from the four universities in the NCR dataset (representing Web of Science) and 63,989 research publications from the same universities in the NSI dataset (representing all peer-reviewed publications in all fields). Our study is work in progress, and the three datasets still need to be matched. Before the deadline for ISSI submissions, we can only briefly describe our methods, present explorative results and conclude preliminarily.

Methods

Our study shares the general research question concerning the relation between publishing productivity and citation impact as studied by Larivière & Costas (2016) and Sandström & van den Besselaar (2016). Our data collection method is different, as explained above. Regarding the two Norwegian data sets on researchers, we will use a similar methodology as was used by Aknes et al. (2013). Within the data set representing WoS (NCR), we will delimit data and choose indicators that make our general findings directly comparable to the two former studies.

For the preliminary results presented below, the data collected by Zhang et al. (2017) is used. The data are derived from an earlier version of NCR (see the three data sets described above) and thereby only represents WoS, not the two other data sets. However, instead of author disambiguation, all researchers have been identified also in the NSI data mentioned above. Specifically, we adopt the following indicators for the preliminary study:

- **Number of Citations (NC)**: the number of received citations in a three-year citation window, starting from the publication year of an article.
• **Relative Citedness (RelCit):** This is a relative citation measure. According to the information provided by Clarivate Analytics, the supplier of NCR, the Relative Citedness of a document is calculated by dividing the actual count of citing items by the expected number of received citations for documents of the same document type (article, in our case), year of publication and subject area. When a document is assigned to more than one subject area an average of the ratios of the actual to expected citations is used. Here the term “subject area” refers to the WoS journal category of the article. For articles published in multidisciplinary journals, NCR uses a paper per paper assignment based on reference analysis.

• **Percentile of Citedness:** Percentile in which the paper falls in its category based on total citations. Indicates that the paper is in the top n% of papers in that category in that publication year - if a document is in more than one category, this is the percentile from the best performing category. We use the top 1% and 10% most cited articles as indicators, as is also done in the two former studies that we relate to and in most other studies.

• **CSS (Characteristic Scores and Scales) Classes:** “a parameter-free solution for the assessment for outstanding performance” firstly developed by Glänzel and Schubert (1988), which can be used to assign single publications in a field and publication year to meaningful impact groups. Characteristic scores are obtained from iteratively truncating samples at their mean value and recalculating the mean of the truncated sample until the procedure is stopped or no new scores are generated. According to Glänzel (2016), all papers could be classified into four groups: “poorly cited” (Class 1), “fairly cited” (Class 2), “remarkably cited” (Class 3) and “outstandingly cited” (Class 4).

**Preliminary explorative results**

Since the three data sets remain to be matched, we will only briefly present some explorative results based on a data sample comprising 899 Norwegian scientists active in the biomedical sciences, following the same data strategy as used in our previous study (Zhang et al., 2016). Our data set is thereby more homogeneous regarding field-specific productivity and publishing patterns than in the two former studies. Only records of the document type “Article” are included and the total number of unique publications (1992-2013) is 18,280. Their citations have been counted until the end of 2015.

As a first step, we calculated for each of different productivity classes the average number of highly cited papers (based on different citedness indicators). Figure 1 shows the relation between productivity and authoring top cited papers (using “outstandingly cited” papers in CSS classes as an example, and results based on other indicators have shown similar profiles). In general, the higher the productivity of a researcher, the more top-cited papers a researcher has.

![Figure 1. Average number of “outstandingly cited” papers (CSS) by productivity class](image-url)
Figure 2 presents the relation between top cited papers and total number of publications. As observed, the lower the citation threshold of the top cited papers, the higher the correlation.

In a second step, to concentrate on principle investigators, we removed the incidental co-authors with low number of publications. Only 564 biomedical scientists with at least 10 articles in the data sample are kept for a further analysis. We ranked the 564 researchers according to their average citedness respectively based on “number of citations (NC)” and “Relative Citedness”, evenly divided into four groups, hence each including 141 researchers. Group 1 stands for the most cited researchers, and Group 4 represents the researchers with least citedness. Although the groups are not exactly the same for the two different citedness measures, we do not find a systematical divergence between the two cases. We then calculated the average value of “total number of articles” for individual researchers in each author group, and the results are shown in Table 1. Interestingly, researchers in Group 1 are much more productive than researchers in Group 4 on average. A striking difference can be observed when Relative Citedness is used for grouping the researchers: Researchers in group 1 are twice as productive as researchers in group 4 on average. The researchers with highest citation visibility are also the most active researchers, and the least cited researchers are found to have the least productivity. However, this is a preliminary finding, and we do not yet know to what degree these researchers are comparable with regard to their backgrounds, positions, resources and roles in researches.

Table 1. The average value of ‘total number of articles’ for individual researchers in different groups
[Data sourced from Thomson Reuters Web of Science Core Collection]

<table>
<thead>
<tr>
<th>Article Citedness</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Citedness</td>
<td>67.95</td>
<td>51.20</td>
<td>49.09</td>
<td>33.81</td>
</tr>
<tr>
<td>Number of Citations</td>
<td>59.70</td>
<td>52.40</td>
<td>53.90</td>
<td>36.10</td>
</tr>
</tbody>
</table>

Our preliminary results are based on data representing the broad field of biomedicine, however with field normalization of citation indicators based on WoS journal categories. In our final study, we will combine this method, applied on all fields in WoS, with methods by which we use the matched data on the researchers and their complete publication records to field.
normalize also the expected productivity in WoS and to study also the factors influencing the productivity by being able to regroup the researchers also according to co-authorship practices, age, gender, position, other output, etc.

Conclusions (main hypotheses)

1. We expect to find similar correlations between publishing productivity and citation impact, as measurable within WoS with bibliometric indicators, as was found by Larivière & and Costas (2016) and Sandström & van den Besselaar (2016).
2. However, we also expect to establish a firmer basis for understanding these observable correlations by studying persons, not just authors, in relation to several factors that may influence productivity differences. Roles, positions, gender, resources, achievements, incentives and bibliometric performances in research are probably interdependent in a complex way and not separable as an interpreted causality (based on correlations) from which we can conclude on the possible effects of evaluation or funding regimes.
3. We find it interesting that the two former studies use a very similar empirical observation to conclude differently about the policy implications for evaluation and funding regimes. We instead expect to be able to enlighten the ongoing debate on the possible effects of such regimes by introducing a methodology not yet used in the several recent studies of the possible effects of the so-called Norwegian model.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (Grant No: 71573085), the FORINNPOL programme (Project No: 256223) at the Research Council of Norway, and the Innovation Talents of Science and Technology in HeNan Province (Grant NO: 16HASTIT038).

References

doi:10.1371/journal.pone.0174205.
On the citation gap of articles naming countries

Philippe Mongeon1, Adèle Paul-Hus2, Antoine Archambaut3, Sarah Cameron-Pesant4, Maxime Sainte-Marie5, Vincent Larivière6

1philippe.mongeon@umontreal.ca 2adele.paulhus@umontreal.ca 3antoine.archambault@umontreal.ca 4sarah.cameron-pesant@umontreal.ca 5maxime.sainte-marie@umontreal.ca 6vincent.lariviere@umontreal.ca

1-6 École de bibliothéconomie et des sciences de l’information, Université de Montréal, Montréal (Canada)
6 Observatoire des sciences et des Technologies (OST), Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Montréal (Canada)

Abstract
Titles of scholarly articles are generally a reflection of their content. They inform the reader about the methods, design, results or conclusion of the study, as well as on the context of the research. The mention of the name of a country, for example, provides a geographical contextualization of the article. In order to better understand the effect of these signaling devices on the reception of a study, this research in progress paper investigates the difference in citation rates of articles that mention a country in their title or abstract and articles that do not. It shows, using WoS-indexed papers published between 1996 and 2013, that mentioning a country in either the title or the abstract is associated with lower citation rates, and that this is observed for every country when all disciplines are combined. The gap in citation rates is also greater in Social Sciences than in other disciplines, which is likely due to their stronger focus on national issues.

Conference Topic
Research on local topics and issues, citation analysis, country-level studies

Introduction
The title of a research article play a central role in the scholarly communication process: it is usually the first point of contact with its potential readers (Diener, 1984; Gross, Harmon & Reidy, 2002; Salager-Meyer & Alcaraz Ariza, 2013). According to Yitzhaki (1994, p. 321), the “primary functions [of a title] are to draw a reader’s attention to a paper and to indicate its content from a short glimpse, thus contributing to its initial selection or rejection”. The title can be considered as the most important summary of a research article since it is “generally the first (and sometimes the only) information obtained from the published article” (Paiva, Lima, & Paiva, 2012, p. 509). It can also reflect the content of a paper by describing its methods, design, results or conclusion, or by revealing important contextual attributes of a piece of research. Similarly, we can posit that mentioning a country name in the title of an article constitutes a geographical contextualization of the article in question.

Several bibliometric studies have investigated the relationship between title characteristics and scholarly impact. Many studies found that articles with shorter titles were more cited (Gnewuch & Wohlrabe, 2017; Paiva et al., 2012; Subotic & Mukherjee, 2014), while others found the opposite (Habibzadeh & Yadollahie, 2010; Jacques & Sebire, 2010). Other studies found that the direction of the correlation between title length and citations differs between disciplines (van Wesel, Wyatt, & ten Haaf, 2014) or found no significant correlation (Rostami, Mohammadpoorasl, & Hajizadeh, 2014). Buter and van Raan (2011) found that the occurrence of at least one non-alphanumeric character in the title was associated with a higher impact, but the correlation strength varied by discipline and was in some cases non-significant or negative.
Titles containing a hyphen or colon (Jacques & Sebire, 2010; Rostami et al., 2014), and titles containing acronyms have also been associated with higher citations rates. According to Subotic and Mukherjee (2014), we should remain cautious when interpreting the results of these studies since “we are only beginning to understand how relevant title characteristics relate to each other in an integrative context” (p. 123). Indeed, “what is said probably matters more than how it is said” (p. 121).

We found five studies investigating the relationship between citations and the presence of a specific geographical region in articles’ title, which is the focus of the present study. Nair and Gibbert (2016) showed that context attributes of titles, such as reference to a specific country, company or industry name, had no significant influence on citations, yet the results of the four other studies point in the opposite direction. Indeed, titles referring to geographical feature were associated with fewer citations by Paiva et al. (2012), Rostami et al. (2014), Jacques and Sebire (2010) and Abramo, D’Angelo and Di Costa (2016). While most of these studies obtained a negative correlation between the geographical contextualization of a research and its scientific impact, they either used small samples (a few hundred publications) or focused on specific disciplines. The study by Abramo et al. (2016) attempted to overcome these limits by analyzing all WoS-indexed Italian publications published between 2004 and 2011. The authors divided articles in a “Country” group containing all publications mentioning “Italy” in their title, abstract or keywords (n=40,024), and a “No country” group including all the remaining publications (n=416,686). Despite being the largest study on the topic performed so far, its methodological design has serious shortcomings. For instance, factors other than the discipline that might affect the number of citations received by a paper were not controlled. Moreover, the “No country” group contains all publications that do not mention Italy, and therefore likely contains articles that mention other countries. Consequently, it is unclear how well the study isolates the effect of country mentions on citations. Another limitation of the research relates to its small geographical scope, both at the level of the researchers (the analysis is restricted to publications by authors affiliated to Italian institutions) and at the level of the research object (the analysis is restricted to publications mentioning Italy). This raises questions regarding the generalizability of their findings to other geographical and thematic contexts.

The present research aims to overcome these shortcomings by analyzing a large dataset of worldwide publication and controlling for article similarity. We also investigate how the amplitude of the citation (dis)advantage of articles naming countries in their title or abstract varies according to the country mentioned. More specifically, we provide answers to the following research questions:

1. What is the relation between the mention of a country in the title or abstract of articles and their relative number of citations?
2. How does this relation differ by discipline?
3. How does this relation differ by mentioned country?

Methods
We used all articles, notes and reviews published in Clarivate Analytics’ Web of Science (WoS) between 1996 and 2013, each assigned to one of four broad categories—Arts and Humanities (AH), Biomedical Research (BM), Natural Sciences and Engineering (NSE) and Social Sciences (SS)—based on the journal’s NSF classification. The twenty countries with the highest number of papers in the WoS over that period were chosen for this analysis (see Table 1). We searched for these countries’ names as well as their associated adjectives and demonyms to
create our “Country” dataset. For each article included in the “Country” dataset, we searched for the most similar article published in the same journal and year, but without any country mention (i.e., not limited to the twenty countries included in the analysis) in the title or abstract\(^1\). To measure the similarity between two articles, we computed the cosine similarity of their respective title and abstract (reduced to their constitutive noun phrases) as well as the Jaccard similarity of their references list (references to WoS source items only). We then ranked candidate pairs per these two criteria and, to ensure maximal proximity between the two articles, only cases where the same article was ranked first per both criteria were considered in the comparative analysis. Table 1 reports the resulting number of article pairs for each country and discipline. The group “Title” includes articles for which the country was mentioned in the title, and the “Abstract” group refers to articles for which the country was mentioned in the abstract, and did not appear in the title. Also, we decided to exclude Arts and Humanities from the analysis since the number of papers in the dataset was too small (a total of 1,456 and 2,161 articles in the “Title” and “Abstract” groups, respectively).

### Table 1. Number of articles with a country name in the title or abstract by country and discipline

<table>
<thead>
<tr>
<th>Country</th>
<th>BM Title</th>
<th>BM Abstract</th>
<th>NSE Title</th>
<th>NSE Abstract</th>
<th>SS Title</th>
<th>SS Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>5,307</td>
<td>10,579</td>
<td>9,118</td>
<td>12,524</td>
<td>2,884</td>
<td>5,971</td>
</tr>
<tr>
<td>Belgium</td>
<td>770</td>
<td>1,837</td>
<td>577</td>
<td>1,551</td>
<td>327</td>
<td>977</td>
</tr>
<tr>
<td>Brazil</td>
<td>6,353</td>
<td>7,555</td>
<td>7,002</td>
<td>9,261</td>
<td>1,539</td>
<td>2,033</td>
</tr>
<tr>
<td>Canada</td>
<td>4,233</td>
<td>9,721</td>
<td>6,969</td>
<td>11,166</td>
<td>2,282</td>
<td>5,247</td>
</tr>
<tr>
<td>China</td>
<td>18,645</td>
<td>21,766</td>
<td>19,608</td>
<td>20,914</td>
<td>6,434</td>
<td>6,845</td>
</tr>
<tr>
<td>France</td>
<td>4,469</td>
<td>9,430</td>
<td>2,931</td>
<td>7,927</td>
<td>1,433</td>
<td>3,164</td>
</tr>
<tr>
<td>Germany</td>
<td>4,470</td>
<td>12,213</td>
<td>2,306</td>
<td>8,337</td>
<td>2,486</td>
<td>5,638</td>
</tr>
<tr>
<td>India</td>
<td>7,134</td>
<td>8,498</td>
<td>8,747</td>
<td>10,746</td>
<td>2,148</td>
<td>2,458</td>
</tr>
<tr>
<td>Italie</td>
<td>5,180</td>
<td>8,339</td>
<td>3,818</td>
<td>9,075</td>
<td>1,555</td>
<td>2,726</td>
</tr>
<tr>
<td>Japan</td>
<td>13,913</td>
<td>15,283</td>
<td>7,977</td>
<td>10,399</td>
<td>2,588</td>
<td>2,984</td>
</tr>
<tr>
<td>Netherlands</td>
<td>3,170</td>
<td>7,383</td>
<td>975</td>
<td>2,659</td>
<td>1,870</td>
<td>4,408</td>
</tr>
<tr>
<td>Poland</td>
<td>1,117</td>
<td>1,719</td>
<td>1,251</td>
<td>2,408</td>
<td>364</td>
<td>676</td>
</tr>
<tr>
<td>Russia</td>
<td>785</td>
<td>1,495</td>
<td>1,500</td>
<td>3,019</td>
<td>639</td>
<td>995</td>
</tr>
<tr>
<td>South Korea</td>
<td>5,772</td>
<td>4,332</td>
<td>3,039</td>
<td>2,914</td>
<td>2,076</td>
<td>1,340</td>
</tr>
<tr>
<td>Spain</td>
<td>4,328</td>
<td>7,889</td>
<td>3,880</td>
<td>9,273</td>
<td>2,385</td>
<td>4,406</td>
</tr>
<tr>
<td>Sweden</td>
<td>3,901</td>
<td>6,841</td>
<td>1,792</td>
<td>3,248</td>
<td>1,623</td>
<td>3,042</td>
</tr>
<tr>
<td>Switzerland</td>
<td>1,501</td>
<td>4,165</td>
<td>1,404</td>
<td>3,152</td>
<td>452</td>
<td>1,059</td>
</tr>
<tr>
<td>Taiwan</td>
<td>3,258</td>
<td>3,621</td>
<td>2,217</td>
<td>2,888</td>
<td>1,873</td>
<td>3,001</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>8,658</td>
<td>26,984</td>
<td>8,238</td>
<td>13,686</td>
<td>7,355</td>
<td>16,827</td>
</tr>
<tr>
<td>United States</td>
<td>11,226</td>
<td>56,393</td>
<td>12,538</td>
<td>80,999</td>
<td>8,153</td>
<td>24,387</td>
</tr>
<tr>
<td>Total</td>
<td>114,190</td>
<td>226,043</td>
<td>105,614</td>
<td>226,146</td>
<td>50,466</td>
<td>98,184</td>
</tr>
</tbody>
</table>

### Results

Figure 1 shows the average difference between the field-normalized citations rates (to enable the comparison of effect size between disciplines) of articles that mention a country in their abstract and their associated control article (i.e., the most similar article with no mention of a country). Four observations can be made. First, the results show that, on average, there is a

\(^1\) To identify mentions of other countries, we used a list of 593 country names, adjectives and demonyms, retrieved from: [https://en.wikipedia.org/wiki/List_of_adjectival_and_demonymic_forms_for_countries_and_nations](https://en.wikipedia.org/wiki/List_of_adjectival_and_demonymic_forms_for_countries_and_nations)
citation disadvantage for articles who mention a country either in their title or abstract. This is true for all countries when all disciplines are combined. Second, the citation disadvantage appears to be smaller when a country is mentioned in the abstract only, except for articles in BM that mention the United States or Germany or Sweden, and articles in NSE mentioning China or Sweden. Third, the citation disadvantage is generally higher in SS than in BM and NSE. Fourth, the citation disadvantage appears to be much higher for some countries than others. While this varies by discipline, we observe that the citation disadvantage is generally smaller for countries who have traditionally dominated the scientific front. There are however notable exceptions, such as papers in NSE mentioning China that appear to have the smallest citation disadvantage. In BM, articles mentioning the United States or Germany in their title also stand out, as they exhibit the opposite results than other countries and fields: they receive on average more citations than the control articles, and they are also more cited on average than articles who mentioned the country only in the abstract.

![Figure 1. Citation (dis)advantage when mentioning a country in the title or abstract by country and discipline.](image)

**Discussion and conclusion**

Our results show that articles mentioning a country in their title or abstract receive on average less citations than similar articles that do not mention a country and, thus, are coherent with those of Paiva et al. (2012), Rostami et al. (2014), Jacques and Sebire (2010) and Abramo et al. (2016). However, since we compared articles from the first group with their closest neighbour (based on the title, abstract and reference similarity), our method allows a better control for the research topic, and thus better isolates the “country effect” than previously used methods. Despite the unequivocal nature of the results, important nuances should be considered in their interpretation. For instance, while the appearance of a country in the title or abstract is likely to reflect the local relevance of an article—or at least its research setting—it is unclear to what extent the absence of geographical contextualization reflects its global relevance.
Consequently, it remains unclear whether the observed citation disadvantage is caused by a perceived lack of global relevance in the eyes of other researchers or by an actual lack of relevance of the articles for research in other geographical or global contexts. The greater citation gap—observed in Social Sciences—can be explained by the fact that research in these disciplines is more often contextualized within a specific region or nation, compared with Natural Sciences where research objects are often by definition international (Gingras & Mosbah-Natanson, 2010). Thus, it could be argued that in Social Sciences, mentioning a country in the title or abstract does act as the marker of a geographical or national context of the research, and that the results may in fact be difficult to transfer to other contexts. This raises a few questions. Is the citation gap caused by the title of the article, or by its content? Would removing the geographical location in title of articles with a local or national focus result in higher citations? Given that the title of a paper serves as a signalling device for those to whom it may be most useful, and that research on local topics is particularly relevant for local stakeholders (Stremersch, Verniers, & Verhoef, 2007), removing markers of geographical context from titles in an attempt to increase citations might come at the expense of other forms of impact.

This research in progress paper also shows that the amplitude of the citation disadvantage varies by mentioned country. Since some countries have a larger research output than others, this may be partially explained by the idea that researchers in a given country are perhaps more likely to do research on local topics, but also more likely to perceive articles mentioning their country as relevant. Moreover, given the importance of citations in research evaluation processes, even researchers in smaller and/or non-English speaking countries might be incentivized to work on research of local issues in larger English-speaking countries like the United States to get more citations (Larivière, 2014). This potential effect on citations might also be amplified by the overrepresentation of research published in the English language in the WoS database (Mongeon & Paul-Hus, 2016).

Further steps of this research will be to 1) extend our list of geographical markers to ensure that the “No country” group does not mention any location. It will also include the country of the authors in the analysis since research on local issues performed by local scientists may be of more local (and less global) relevance than topics that are studied by the global scientific community, which might also influence citation rates. Finally, there is empirical evidence that research focusing on local issues tends to be more interdisciplinary (Chavarro, Tang, & Rafols, 2014), and that interdisciplinarity is often associated with higher citation rates (Larivière, Haustein, & Börner, 2015). Thus, further research should also look at how geographical contextualisation of research is related to interdisciplinarity, and how this may affect citation rates.

Acknowledgements
This study was funded by the Social Sciences and Humanities Research Council of Canada and the Canada Research Chair on the Transformation of Scholarly Communications.

References
of Informetrics, 5(4), 608-617.
Identifying Funding Allocation Gaps Within an Academic Domain: Case Study of Robotics Research

Cristian Mejia 1 Yuya Kajikawa 2

1 mejia.c.aa@m.titech.ac.jp
Tokyo Institute of Technology (Japan)

2 kajikawa@mot.titech.ac.jp
Tokyo Institute of Technology (Japan)

Abstract
In this paper, we compare the citation network of academic articles acknowledging financial funding, to the network of articles in the same research domain without funding acknowledgement. We found similar clusters in both networks, although with differences in publication intensity and academic impact. Other clusters were exclusive of any of the two networks, those only found in the network without financial acknowledgements represent funding allocation gaps. We applied our method to identify gaps in robotics research where we found topics like micro-robotics, spherical robots, and cloud-computing for robotics being underfunded. The proposed method may benefit two groups, researcher can identify topics receiving more attention, thus, having higher possibility of getting funds; and policy makers may monitor funding-independent clusters, born of researchers own curiosity, in order to consider those topics in future funding strategies.

Conference Topic
Methods and Techniques; Citation and co-citation analysis; Research funding

Introduction
The study of research funding is a growing field in bibliometrics. Large academic data bases like the Web of Science and Scopus now include funding information extracted from the acknowledgement section of the articles, opening up the possibilities for understanding the trends of funding across different countries or disciplines (Wang & Shapira, 2011). However, most of the research is focused on understanding the role of funding in relation to positive academic impact, usually measured by citations received by the funded projects (Gök, Rigby, & Shapira, 2016). While just few have pointed out the need of using maps of funding sources to understand gaps and new research trends (Rigby, 2011).

We take further steps in that direction, by trying to identify missing spots in the allocation strategies of funding organizations. To do so, we draw upon the literature on finding missing connections or common patterns by the comparison of two or more modes of knowledge representation. For instance, Shibata et al (2010) extracted commercialization gaps by comparing clusters in a citation network of patents with those found in a network of papers for the case of solar cells. Active clusters in the paper network, not linked to any cluster in the patent network were described as commercialization opportunity. Similar linkage strategy was followed by Ittipanuvat et al (2014) when connecting two different research domains, robotics and gerontology. While the focus in that research was not about missing connections, text similarity measurements proven to be able to stablish connections between two fields that are semantically different.

In this article, we take a different approach by adapting those methodologies within the same domain and type of documents. Although, this time two modes of the same type of documents are compared: the set of articles reporting financial funding and those which do not. We aim to provide a methodology for the identification of specific research topics not receiving funding,
and explore the academic landscapes created by the citation networks of funded and non-funded articles. We applied the methodology to the robotics research, which is known for its disciplinary breadth and funding support.

The rest of the paper is composed as follows. We describe the data extraction strategy and methodological treatment in the data and methods section. Characteristics of the data and findings derived from the citation network and text linkage is shown and discussed next. We conclude by summarizing those findings.

Data and Methods

An overview of the proposed methodology is represented in figure 1. Data were retrieved from the Web of Science. We searched for articles on the topic “robot” from 2009 to 2016, obtaining 76,209 records. The selection of 2009 as starting year was determined by the availability of funding acknowledgements in the database (Paul-Hus, Desrochers, & Costas, 2016). We split the records in two parts, those mentioning at least one funding organization, and those without funding acknowledgment. The former accounted for 21,142 records (27.7%), and the latter for 55,067 (72.3%). The two datasets were normalized and compared in terms of year frequency, participating countries, and target research categories.

Each of these data sets received the same following treatment. We created a direct citation network (de Solla Price, 1965) where the articles are treated as nodes, connected to other nodes according to the cited references found in each article. This type of networks is known for generating accurate taxonomies of scientific knowledge (Klavans & Boyack, 2017), and for detecting research fronts (Shibata, Kajikawa, Takeda, & Matsushima, 2009). Direct citation networks have been applied to map global scientific landscapes, and specific research domains. Once the network is created, the direction of the citations is neglected and the largest connected component extracted.

Following, we identified tightly connected communities of nodes by applying a non-fuzzy, topological algorithm based on modularity maximization (Newman & Girvan, 2004). This algorithm looks for the best partition by contrasting the network to a random graph generated from the features of the original one, thus, the number of clusters is automatically defined by the natural structure of the network. Then, we used the large graph layout (Adai, Date, Wieland, & Marcotte, 2004) to visualize the network. The size of nodes was set to zero, and clusters are depicted by the edges only. At this point we have two networks, one representing the knowledge derived from articles that acknowledge funding, and one without funding acknowledgement. Finally, we compared the communities in both networks in term of text similarity. For each cluster, we obtained the set of keywords that best represent them by weighting them using the term-frequency/inverse-document-frequency adapted to cluster level (i.e. inverse-cluster-
frequency), those keywords were translated to a vector space model and the cosine similarity was computed. Cosine similarity ranges from 0 to 1, being 1 the exact match. We regarded clusters to be similar if they score a similarity of 0.5 or more.

Once similar clusters were connected, we identified three types of clusters: a) Clusters that exist in both networks; b) clusters exclusive to the network with funding acknowledgements; and c) Clusters exclusive of the network without funding acknowledgement. The third type represents the funding allocation gap. In this article, allocation gaps are defined from a bibliometric standpoint only, being the clusters of knowledge not receiving funding. In practice, their interpretation varies from those areas overlooked by funding agencies, to ideas explored without the need of grants.

**Results and discussion**

Academic articles on robotics research were obtained and classified according to the funding information provided. 27.7% acknowledge at least one funding organization. This proportion is in line with the general proportion of articles reporting funding in the Web of Science’ Core Collection (Paul-Hus et al., 2016). After normalizing the data, we compared them in terms of year frequency, and participating countries as shown in Figure 2. Robotics is an ever-growing field. Before 2012 shares of articles reporting, or not, funding were balanced. However, in recent years the difference is getting noticeable. 2016 has the largest share of articles reporting funds, while 2015 is the largest for articles without acknowledgement. In the case of participating countries, USA and China are the largest contributors. Some countries like China and South Korea have larger shares when it comes to acknowledged funding, while Iran and India show an opposite behaviour. England and Australia have similar proportions in both scenarios.

![Figure 2](image.png)

**Figure 2. (a) Normalized frequency per year (b) Top 15 participating countries.**

There are differences regarding specific research categories within robotics. Figure 3 depicts such difference in an overlay map (Rafols, Porter, & Leydesdorff, 2010), plotted using VOSviewer (Van Eck & Waltman, 2010).

While surgery is present in both sets, there is a clear intensification of funding in robotics for medical sciences, particularly neuroscience and rehabilitation. Multidisciplinary science is also benefited from funding, and so is intensified the region of applied sciences. The social sciences and humanities are underrepresented.
Figure 3. Target research categories within robotics for articles (a) not reporting funding, and (b) reporting funding.

Figure 4 shows the citation networks for the two data sets. Clusters were labelled by inspecting the relevant keywords, and the title and abstract of the most cited papers within them. We compared the clusters by using cosine similarity and found 17 highly similar pairs of clusters, which top ten are characterized in table 1. Clusters from the network without funding acknowledgement have an average year that is younger than their counterparts in 13 of the cases, although by a small difference. On the other hand, the performance measured in terms of cites received per articles within each cluster, is clearly better for the network reporting funding, except for the most similar connection about the topic of transoral robotic surgery.

The remaining clusters were found of little similarity between the two networks. However, several clusters were of small size, so that we focused only in the largest clusters that accounts for the 80% of articles in each network. Exclusive clusters are shown in table 2. Again, clusters from the network reporting funding show better performance in terms of citations received. By exploring the robotics research separately, we could find similarities and gaps between the two networks structures, however the two sets of articles do not exist in isolation. Articles with and without funding cite each other forming the complete academic landscape of a domain. A simpler method to evaluate the presence of funding is to create a citation network of all papers, obtain the clusters, and compute the proportion of articles reporting funding within each cluster, expecting to find some of them ranging from purely funded to not funded at all. That practical and straightforward method may serve as indicator of funding; however, it leads to a different interpretation to that aimed in this article. We created the citation network by using all the data on robotics, obtaining a network split in 460 clusters. 18 of them aggregated 90% of the articles.
The proportion of articles acknowledging funding within each of them was rather stable, having a mean 33.3%. This suggests that the specific topics in robotics are receiving funding in a proportion close to the global average. Moreover, we did not find any relevant cluster without presence of funding. Finally, we compared the 4 clusters without funding shown in Table 2, to the clusters in the merged network. Articles from the micro-robotics cluster were found in the merged network, forming a new cluster along with Nano-robotics having 42% of funding presence. Articles from the other 3 clusters were scattered into several clusters in the merged network. Therefore, our proposed methodology generates clusters that are fine grained, and can inform about network structures that are independent of funding.

Table 1. Correspondence of clusters between the networks of articles reporting (F) and not reporting (NF) funding

<table>
<thead>
<tr>
<th>Topic ID*</th>
<th>Cluster ID</th>
<th>Sim.</th>
<th>Average year</th>
<th>Size (%)</th>
<th>Cites per paper</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>NF</td>
<td></td>
<td>F</td>
<td>NF</td>
</tr>
<tr>
<td>1</td>
<td>22</td>
<td>6</td>
<td>0.89</td>
<td>2014.5</td>
<td>2013.3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>5</td>
<td>0.88</td>
<td>2013.4</td>
<td>2013.1</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>2</td>
<td>0.83</td>
<td>2013.5</td>
<td>2012.9</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>8</td>
<td>0.78</td>
<td>2013.2</td>
<td>2013.0</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>11</td>
<td>0.74</td>
<td>2012.8</td>
<td>2013.0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>4</td>
<td>0.70</td>
<td>2013.3</td>
<td>2013.0</td>
</tr>
<tr>
<td>7</td>
<td>27</td>
<td>22</td>
<td>0.65</td>
<td>2013.4</td>
<td>2012.9</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>28</td>
<td>0.65</td>
<td>2013.3</td>
<td>2013.1</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>3</td>
<td>0.65</td>
<td>2013.4</td>
<td>2013.5</td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>40</td>
<td>0.64</td>
<td>2013.3</td>
<td>2012.9</td>
</tr>
</tbody>
</table>

* 1-Transoral robotic surgery; 2-Exoskeleton and prosthetics for rehabilitation; 3-Robot-assisted prostatectomy; 4-Parallel manipulators; 5-Biped walking; 6-Algorithms for motion planning and control; 7-Mobile robot gathering; 8-Space robotics; 9-Social robotics; 10-Hand dexterity: Grasping

Finally, challenges exist when using acknowledgements as proxy of funding. The proposed methodology depends on the reliability of rightly acknowledging the funding sources, and this presents two limitations. First, authors willingly or unwillingly may omit an acknowledgement. The reasons may vary from cultural and political as discussed in Rigby (2011), or because some funding bodies, private specially, prefer to be omitted as mentioned by Wang and Shapira.
Second, data providers as the Web of Science may fail to rightly recognize the entities written in the acknowledgement section when using automated tools to populate the article metadata. In regard of these limitations, experiments realized in the field of cancer research, a field known for its intense funding activity, have demonstrated that less than 3% of papers in the corpus omitted any of the sponsors involved in the research; and that the Web of Science reach a precision and recall above 90% when converting the acknowledgment text to funding metadata (Grassano, Rotolo, Hutton, Lang, & Hopkins, 2016). Therefore, the use of acknowledgement can be associated to patterns of funding, particularly for the study of general trends as proposed in this article.

Conclusion

We took advantage of the financial acknowledgement information provided in academic articles to separate the corpus of research reporting and not reporting funding. From these sets of records, we build citation networks that allowed us to explore the academic landscape of funded and unfunded research. The comparison of these networks leaded us to three categories. The first one, specific topics that exist in both networks, as is the case of robotics for medical science, particularly related to robotic surgery and rehabilitation, this group also brought to surface the relation between funding and academic impact. Except for research on transoral surgery, clusters with common topics had better citation per paper ratio in the network of funded articles, this result agrees with the bibliometric literature arguing that the presence of funding leads to more citations (Lewison & Dawson, 1998). The second group, are clusters only present in the network of funded articles, here we found swarm robotics, temporal logic and neural network algorithms, and robots for harvesting. The final group are those clusters that appeared in the network of articles without funding acknowledgements, thus, representing funding allocation gaps. These underfunded topics were micro-robotics, spherical robots, and cloud computing for robotics. It also appeared a cluster of general concepts about robotics technologies. The methodology proposed in this article allows the identification of topics receiving high or little attention from funding organizations. The study of landscapes of research funding may facilitate the creation of research strategies, by understanding the trends of funding. However, the assessment of relevance of the funding gaps, or its typology (e.g. basic vs applied, breakthrough vs conventional) is still subject of future research.

By applying the proposed method researchers get a picture of specific topics in an academic domain that are receiving higher attention and attracting funding, while allows them to avoid those with little possibilities of getting grants.

Acknowledgments

This research is partially supported by a scholarship from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (A) 26245044.

References


A Logical Comparison of Citation Wakes and Generations

David Friggens

david.friggens@mbie.govt.nz
Ministry of Business, Innovation and Employment, Wellington (New Zealand)

Abstract
We provide an empirical comparison of two indicators – the citation-wake-score and the $fp^k$-index – for measuring an article’s influence by counting all of its direct and indirect citations, i.e. its citations, citations of citations, and so forth. The indicators are compared for the small field of mathematical logic, but articles from all fields in the Scopus database are used for the calculations. We find that the two indicators are very similar. Furthermore, the use of a general bibliometric database raises different questions for future investigation than previous results using field specific data sets.

Conference Topic
Indicators; Citation and co-citation analysis

Introduction
Most citation-based indicators in bibliometrics consider only direct citations between articles (Waltman, 2016), though there are approaches that consider indirect citations (Fragkiadaki & Evangelidis, 2014). Recently, Klosik & Bornholdt (2014) and Fragkiadaki & Evangelidis (2016) have independently proposed new indicators that count all of the direct and indirect citations of an article, i.e. an article’s citations, citations of citations, citations of citations of citations, and so on, each level being a new generation. The collective generations of citations form an article’s citation wake. Both indicators are similar, with differences in the ways they weight the age of an article and the generation of a citation. Additionally, the choice of a manual parameter differs too: the first parameterises the generation weighting and the second limits the number of generations counted.

In this study we provide two contributions: first by comparing the two indicators, and second by providing new empirical data. Both previous studies used field-specific data sets, from the physics journal Physical Review and the computer science database DBLP, respectively. We look at one field – mathematical logic – but consider indirect citations from any field in the Scopus database. The choice of a mathematics subfield was in part as it is commonly used as an example of a field with low citation rates (e.g. Waltman, 2016, Section 6); are indirect citations a more important indicator of an article’s influence in a field with low citations?

Indicators
We consider two indicators: the wake-citation-score of Klosik & Bornholdt (2014) and the $fp^k$-index of Fragkiadaki & Evangelidis (2016); we shall refer to them (both indicator and reference) as KB14 and FE16, respectively.

First, let us define some concepts and a generalised indicator. For a given data set of publications and their citation network, let $P(y)$ be the number of publications published in year $y$. For a specific publication $p$, let $G(p, g)$ be size of generation $g$ – the number of publications whose shortest citation path to $p$ is length $g$. Thus, the number of direct citations is $G(p, 1)$.

Here we follow the terminology of FE16. KB14 term “generation $g$” as “neighbourhood layer $g - 1$.”
For a given publication \( p \) published in year \( y_p \), we can define a generalised indicator \( \mathbb{I}(p) \) to assess the impact of \( p \) as the sum of the generation sizes, up to a given limit, and weighted by a damping factor for each generation; a constant is added, and the result is normalized based on its age. Formally, we write:

\[
\mathbb{I}(p) = c + \sum_{g=1}^{\text{lim}} (\text{df}(g) \times \mathbb{G}(p,g)) \over \text{an}(y_p)
\]

The values of the components for each of the two specific indicators are shown in Table 1, and discussed in more detail below. Each indicator has a configurable parameter, so actually defines a family of indicators. For KB14 it is \( 0 \leq \delta \leq 1 \) in the damping factor, for FE16 it is \( \lambda \geq 1 \) to limit the number of generations counted. Note that we have deliberately used different notation from both KB14 and FE16 to avoid confusion and make the commonalities clearer.

### Table 1. Components of the two indicators

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
<th>KB14</th>
<th>FE16</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c )</td>
<td>constant</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( \text{lim} )</td>
<td>generation limit</td>
<td>( \infty )</td>
<td>( \lambda )</td>
</tr>
<tr>
<td>( \text{df}(g) )</td>
<td>damping factor</td>
<td>( \delta(g-1) )</td>
<td>( \frac{1}{g} )</td>
</tr>
<tr>
<td>( \text{an}(y_p) )</td>
<td>age normalization</td>
<td>( \sum_{i=y_p}^{y_{\max}} \mathbb{P}(i) )</td>
<td>( y_{\max} - y_p + 1 )</td>
</tr>
</tbody>
</table>

**Damping Factor**

FE16 note that later generations should count less than earlier ones, and propose that the count should be divided by the generation number. KB14 propose a configurable damping factor, allowing for all generations to be counted equally (\( \delta = 1 \)) through to only counting direct citations (\( \delta = 0 \)); they report experimental results primarily with \( \delta = 0.9 \). Figure 1 compares the two approaches, and shows that the reductions are not comparable.

**Age Normalisation**

Older publications have had more opportunity to receive citations than younger ones, so direct comparisons would be unfair. FE16 simply divide by the age, starting from 1 in the year of...
publication. KB14 divide by the theoretical maximum citation wake possible – the count of all publications in the data set from the year of publication onwards. The latter approach could have different effects based on the data set used, but in the Scopus database between 2002 and 2015 the two approaches are very similar, as shown in Figure 2.

**Generation Limit**

KB14 assume all generations of indirect citations should be counted. FE16 propose limiting the depth, and report experimental results for $\lambda = 3$.\(^2\) In practice, lower values of $\delta$ with KB14 may be equivalent to limiting the generations counted. If we assume that a dampened generation count less than 1 is equivalent to being excluded, then Table 2 shows the values of $\delta$ that match $\lambda$ under different maximum generation sizes.

<table>
<thead>
<tr>
<th>max gen size</th>
<th>$\lambda$</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>30</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>$\delta$</td>
<td>0.100</td>
<td>0.316</td>
<td>0.599</td>
<td>0.720</td>
<td>0.785</td>
<td>0.853</td>
<td>0.889</td>
</tr>
<tr>
<td>1,000</td>
<td>$\delta$</td>
<td>0.032</td>
<td>0.178</td>
<td>0.464</td>
<td>0.611</td>
<td>0.695</td>
<td>0.788</td>
<td>0.838</td>
</tr>
<tr>
<td>10,000</td>
<td>$\delta$</td>
<td>0.010</td>
<td>0.100</td>
<td>0.359</td>
<td>0.518</td>
<td>0.616</td>
<td>0.728</td>
<td>0.790</td>
</tr>
<tr>
<td>100,000</td>
<td>$\delta$</td>
<td>0.003</td>
<td>0.056</td>
<td>0.278</td>
<td>0.439</td>
<td>0.546</td>
<td>0.672</td>
<td>0.744</td>
</tr>
</tbody>
</table>

**Data**

In this study we used a full 14 year extract from the Scopus database (Elsevier, 2016), spanning 2002 to 2015 and containing 24,964,475 publications of type “article” or “conference paper”, with 208,583,913 citations between them. We calculated indicators for the 13,354 publications from sources classified with the ASJC code 2609 (Mathematics; Logic).\(^3\)

In comparison, Klosik & Bornholdt (2014) used a data set spanning 116 years (1893-2009) from a single physics journal (*Physical Review*) containing 438,000 publications and 4.5 million citations. Fragkiadaki & Evangelidis (2016) used a data set from DBLP containing 20,000 computer science publications and roughly the same number of citations.

**Results**

**Generations**

There were 13,354 publications from mathematical logic classified sources, of which 8,367 had citations. This group received direct citations from 34,692 publications, and indirect citations from a further 2,439,601 publications. Taking the logic publications as generation 0, there are 23 more generations, peaking at generations 7 and 8, as seen in Figure 3. The publications come from all 27 ASJC top-level fields. Figure 5 plots the percent of each field across all generations versus the percent in the whole data set. Sources can receive multiple field classifications, so the totals add up to more than 100%.

For individual publications, the maximum generation ranges up to 31, as shown in Figure 4. Most publications have smaller generation ranges: 28% only have 1 generation, 58% have no more than 3 generations, 87% have no more than 8, and 99% have no more than 20

\(^2\)“This number has been chosen based on the authors’ sentiment that three generations (similar to friends of friends of friends in social networks) are enough to illustrate the usability and validity of the indicator under different circumstances.” (Fragkiadaki & Evangelidis, 2016, p670)

\(^3\)To be clear, Scopus classifies the source (e.g. journal) rather than the article. Some included articles may not actually be about logic (if the source covers multiple subfields) and, conversely, some logic articles may not be included, as their source has a broader mathematics classification or is multidisciplinary (e.g. *PLoS ONE*).
generations. The size of each generation per publication is mostly small – 13% contain only one indirect publication – but they range up to nearly 277,000 in one case. 48% have up to 10 publications, 79% up to 100 and 94% up to 1,000.

Figure 3. Size of generations from set of logic publications

Figure 4. Number of generations for each logic publication, darker parts of the bar indicate older publications

Figure 5. Comparison of field distribution in logic generations vs whole data set

**Indicators**

Figure 6 shows comparisons between five indicators: the average citations per year (CPY), the Mean Normalized Citation Score (Waltman et al., 2011), FE16 with $\lambda$ values of 3 (used by the authors) and 31 (the maximum in this data set), and KB14 with $\delta$ of 0.9 (used by the authors). The figure displays Pearson and Spearman correlation coefficients, and a scatter plot to visualise the relationship between each of the ten pairs of indicators.

The most striking result is how similar the two full generation indicators are. Another observation is the L shape of the bottom left scatter plots, which results in low Pearson
correlations. The direct citation indicators favour young publications with relatively high citations that are unlikely to be sustained, but these are not favoured by the new indicators. The limited FE16 indicator appears visually more correlated to the direct indicators, but has a higher rank order correlation with the other two indirect indicators.

![Matrix of pairwise comparisons of five indicators, showing scatter plots, and Pearson and Spearman correlation coefficients](image)

Figure 6. Matrix of pairwise comparisons of five indicators, showing scatter plots, and Pearson and Spearman correlation coefficients

An alternative comparison is to consider the top 10% of publications, ranked by each indicator. Figure 7 contains an UpSet (Lex et al., 2014), an alternative to a Venn diagram, visualising the overlapping rankings. Only 3.3% of the publications were ranked in the top 10% by all five indicators.
Discussion

Our results show that the two indicators proposed independently by Klosik & Bornholdt (2014) and Fragkiadaki & Evangelidis (2016) are quite close, and with similar configuration produce very similar results. Further research is needed to investigate if there are situations where one approach has a clear advantage, but for mathematical logic the differences are much less than they may appear at first glance.

The biggest difference between the two indicators is the choice of parameter: count all generations and tune the damping factor, or place a fixed generational limit? We have shown that there may be equivalent configurations, so the main practical concern may be ease of computation.

Computing these indicators using a broad data set like Scopus raised new questions that remain to be investigated, especially about field normalization. So far the indicators have been used to look at publications from a single field – do we need to divide by expected values for cross field comparison as with MNCS? Even for a small field like mathematical logic the wakes include publications from all fields, so perhaps this is unnecessary? Alternatively, perhaps field normalization needs to take place within the counting of the indirect citations at each generation?

References


A novel analysis on global environmental quality research in the last two decades based on bibliometric modeling indicators

Dron Lha¹, Lihe Chai²

¹ lhakdron0920@163.com
Tianjin University & Tibet Institute of Plateau Atmospheric and Environmental Science (China)

² lhchai@tju.edu.cn
Tianjin University, Tianjin (China)

Abstract
A novel analysis on the global environmental quality research (EQR) during the last two decades is conducted based on bibliometric modeling indicators. Four aspects of findings are obtained. (1) From the subject categories distribution and the co-words network, we elucidate the global EQR involves both natural aspects and social aspects. Small world (c=0.17 and L=3.042) of co-words network highlight the underlying strong relations between most topics. (2) For the cooperative network, a large clustering coefficient (0.367) and a small average path length (1.909) indicate very accessible cooperation between any countries during the EQR. Power-law distribution of node degrees implies that led by USA and China, all over the world is involved in EQR by a self-organized way. However a large slope (α=-3.958) means many countries have not yet sufficiently played their roles. (3) The Zipf’s distribution for citations, Bradford’s distribution for journals and Lotka’s distribution for funds reflect the self-organized information transmission in the global EQR. Bradford’s distribution suggests the participations of wider scopes of journals in EQR are expected. Lotka’s distribution for funds with α=-1.9 implies the funding agencies distribution is too concentrative. (4) By elucidating the underlying pattern of global EQR, some coping strategies are proposed.

Introduction
The interactions between the environment and the activities of all living beings in the world are very strong (Green et al. 2000). During past years, there were extensive environmental quality research (EQR) from global scientists with diverse scopes involved. Some researches were conducted from agricultural perspectives by relating natural resources protection to environmental quality improvement (Cassman et al. 2003), by relating resource use efficiency to environmental quality (Fan et al. 2011), by analyzing the impact of soil erosion on environmental quality (Lal 1998) and the impact of crop residue removal on environmental quality (Blanco-Canqui & Lal 2009), and by aimed at developing an integrated environmental risk assessment and management (IERAM) framework for enhancing the sustainability of such MPAs (Xu et al. 2015). There were also studies about environmental quality and potential evolution from the view of animal populations (Charmantier & Garant 2005; Rose 2000). Brasington & Hite (2005) and Mitchell et al. (2007) conducted EQR from the view of indoor environment in a small angular space. Farzin & Bond (2006) analyzed
democracy and environmental quality to illustrate the relations between social
environmental quality and economical environmental quality. Le et al. (2016)
elaborated the relationship between trade openness and environmental quality. In
addition, there were also lots of researches focusing on other numerous scopes, such
as climate impact of Russia (Frijters & Van Praag 1998), environmental quality in
Taipei (Tzeng et al. 2002) and the local environmental quality or life-satisfaction in
German (Rehdanz & Maddison 2008), and so on.

Doubtless, it is very necessary to uncover the underlying pattern of EQR from a large
scale of literature of EQR. This paper attempts to conduct an intensive bibliometric
modeling indicative anatomy on the pattern of global EQR.

Methods and data sources

Methods

The paper adopts statistical analyses on subject categories distribution to reveal the
connotation of EQR from a macroscopic view. By constructing co-words network and
calculating its dynamic parameters from complex network theory, the connotation of
EQR is revealed from microscopic keywords. In this way, the connotation of EQR
can be strictly explored from the new perspective of complex network by detailed
parameter calculations (e.g. small world phenomena). By constructing cooperation
network of countries and calculating its dynamic parameters from complex network
theory, the global cooperation actions on EQR can be rigorously elucidated by a
detailed calculation of the network parameters (e.g. small world phenomena,
scale-free degree distribution). Some other factors (i.e., citation dynamics, journal
dynamics and funding dynamics) on the global actions of EQR can be analyzed by
utilizing three prime indicators of bibliometrics: Zipf’s distribution for citation,
Bradford’s distribution for journals and Lotka’s expression for funding agencies
distribution. Finally, based on the findings from the bibliometric modeling indicative
results, the corresponding coping strategies can be proposed.

Data sources

The data were retrieved from Web of Science with the search term “environmen*
quality” from the topics, abstracts, and keywords of all publications related to EQR
during 1998-2015. A total of 6819 publications of EQR were obtained. As shown in
Figure 1, journal articles increased year by year, reaching a peak in 2015. Moreover,
proceedings and reviews had the most publications in 2012 and 2015, respectively.

We analyze dynamic characteristics of the data. We consider the empirical Logistic
correlation for the relationship between accumulative publications \(y\) and time \(x\)

\[
y(x) = \frac{k}{1 + ae^{-bx}}. \quad (1)
\]

In Eq. (1), \(k\), \(a\) and \(b\) are coefficients. \(k\) represents the theoretical maximum possible
accumulative publications that EQR could achieve. \(b\) is related to growth rate. That is,
\(b\) relates the time of reaching the theoretical maximal publication accumulation: the
smaller \(b\) corresponds the longer time. By the correlations between the accumulative
publications and time as shown in Figure 2, it is clear that Logistic formula is
generally met, which means the accumulative publication output of EQR holds Logistic growth pattern. That is, generally, accumulative publication output of EQR would undergo a near exponential growth at initial stages, and reach a maximum growth rate at $x=(\ln a)/b$ with $y=k/2$, and then would be gradually close to a stable theoretical value $k$ at a constantly decreasing growth rate. From Figure 2, we $k=8290$, $b=0.23$. The large $k$ and small $b$ well validate that in current phase (1998-present), the EQR is still in the early stages of a rapid development as an emerging discipline. Our data source from the current phase (1998-present), which is characteristic of early and rapid developing movements, is enough for our analyses on the pattern of global EQR, including what is EQR, and how we cooperate to address the issues of global EQR.

![Figure 1. The publication output](image1)

![Figure 2. The Logistic growth of publication output](image2)

**Results and discussions**

*The connotation of environmental quality research (EQR)*

We first uncover “what the connotation is” from the EQR’s emergence and the process of reaching the consensus by analyzing the publications of EQR. This can be done from the distribution of macroscopic subject categories and the evolutionary co-words network of microscopic keywords.

(1) EQR’s connotation from the distribution of macroscopic subject categories

Since each publication can be assigned a subject category, the stable distribution of subject categories that have formed during a long time can reflect what the agreed connotation of a discipline and how the scientists reach the consensus during the emerging movements or efforts of developing a discipline.

In this study, EQR covers 177 subject categories in SCI-E and SSCI. By statistical calculations, a logarithmic scale radar chart is obtained as in Figure 3. We can see EQR covers a wide range of subjects, including environmental science, agriculture, economy, architecture, computer science, transportation, administration, medicine, psychology, etc. The scientists have gradually reached the consensus that the connotation of EQR includes not only the natural aspects (air, water, soil, physics, chemistry, biological life, etc), but the social aspects (economy, culture, psychology, education, etc). Environmental science is dominant in EQR, followed by engineering and agricultural science, economics and architecture, etc. In Figure 4, we analyze the dynamic trends of the top 6 subject categories. Environmental science ecology held a relative stable growth, reaching a peak in 2015. Engineering, agriculture and business
economics sometimes had downward trends, which reached the peaks respectively in 2012, 2008 and 2009. Starting from 2006, construction building technology had a most obvious growth trend. Water resources showed a steady growth from 2010. During past years, EQR flourished in environmental science ecology. Construction building technology had a greater growth than engineering, agriculture and economics. According to the accumulative publications of the top 6 subjects, Logical growths of Eq. (1) are correlated as in Figure 5. A deep understanding on the subjects can be obtained by comparing its parameters $k$ and $b$. Large $k$ and small $b$ mean that the vitality of the subjects in EQR. Environmental science ecology ($k=4869.26$ and $b=0.23$) has a high growth vitality. Agriculture ($k=693.79$ and $b=0.26$) and business economics ($k=658.47$ and $b=0.31$) have a relatively low growth vitality (i.e., easy to be saturated). In EQR, the most flourishing subject seems to be environmental science ecology.

Figure 3. The subjects distribution Figure 4. The annual variations of the top 6 subjects

Beyond qualitative understanding, we here use the detailed data to clarify the connotation of EQR by elucidating the distribution of EQR’s subject categories that have formed during past years. This can partly reflect the agreed connotation of EQR resulted from the emerging movements or efforts in EQR.

(2) EQR’s connotation from microscopic co-words network

The keywords of a publication can accurately interpret the microscopic content of a
With keywords as nodes and co-occurrence of two keywords as the edges between nodes, a co-words network can be constructed. By the analysis of a co-words network, i.e., its topological structure or clustering and networked parameters, we can uncover what connotation of an issue is agreed by scientists, and how scientists gradually reach the consensus on the issue during the emerging movements or efforts of developing a discipline. Thus, besides the distribution of EQR’s subject categories, which provide an understanding on EQR’s connotation from a macro viewpoint, we here further use co-words network to reflect what the agreed connotation is, and how we reach the consensus of EQR’s connotation from a new flavor.

As shown in Figure 6, the co-words network pattern is composed of 998 nodes (keywords) and 7294 edges (connections between keywords). From the classified color areas, the indoor environment, the sustainable environment, water quality and heavy metal are the centers of proliferating domains, suggesting that they dominate the connotation of EQR during past years. As Figure 6, water quality-centered keywords include monitoring, sewage, eutrophication, nutrient, phosphorus, agriculture, land use and management, and so on. With heavy metal as the center, the most closely related keywords are metal, soil, sedimentary pollutants, risk assessment, biological monitoring, water framework indicators, etc. Environment-centered keywords have closest relations with sustainable development, environmental Kuznets curve, indicators, climate change, China, model, geographic information systems, risk, environmental policy, water and air pollution, and so on. The keywords that are closely related to indoor environment include health, indoor air quality, indoor environment, comfort, energy consumption, energy, temperature and schools.

Figure 6. The co-words network pattern of EQR
Doubtless, during past years environmental quality was largely changed by mankind’s activities. These changes even endanger human living and development. The Earth we live together is facing big challenges. By a large number of actual movements and studies on environmental quality from all angles, a lot of publications on EQR appeared. These efforts from all aspects were embodied in the evolutionary keywords (i.e., co-words network). From a new flavor, Figure 6 can partly illustrate what efforts were made, what the agreed connotation is, and how we reach the consensus of EQR’s connotation.

A hotspot in a discipline neither merely consists of a single keyword, nor assembles a group of keywords with poor correlations. The connections of the co-words network are complicated. For example, pollution is related to both heavy metals and water quality or environment, but its connections with the heavy metal are closer. By calculating the parameters of the co-words network, we can learn more quantitatively the connections and then more accurately understand the agreed connotation of EQR. Co-words network is a complex network, we have to calculate the main parameters related to dynamic characteristics, namely average clustering coefficient $c$, average path length $L$, average degree $k$, network diameter $D$ and network density $p$.

The parameters of the co-words network are calculated, as shown in Table 1. For comparisons, we establish random network (with 998 nodes and links probability of 0.146) and calculate its parameters, as also shown in Table 1. From the comparable data, we find the average clustering coefficient of the network is 0.17, much larger than the average clustering coefficient 0.007 of corresponding random network. However, the average path length of the network is 3.042, smaller than that of corresponding random network. These results show that the co-words network exhibits evident small world phenomena (Strogatz 2001). This finding is meaningful. In one way, we can say small world phenomena that widely exist in other types of networks may hold well for co-words network. In other way, this result can help us to better understand the connotation of EQR: We can draw that the EQR discipline is currently at rapid development stage; The relationship between factors within disciplines is very close, holding strong correlation between keywords and the great co-occurrence chance.

Table 1. The calculated parameters for the co-words network of EQR

<table>
<thead>
<tr>
<th>Parameters</th>
<th>The co-words network</th>
<th>The random network</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>998</td>
<td>998</td>
</tr>
<tr>
<td>$M$</td>
<td>7294</td>
<td>7387</td>
</tr>
<tr>
<td>$c$</td>
<td>0.17</td>
<td>0.007</td>
</tr>
<tr>
<td>$L$</td>
<td>3.042</td>
<td>3.596</td>
</tr>
<tr>
<td>$k$</td>
<td>14.617</td>
<td>14.804</td>
</tr>
<tr>
<td>$D$</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>$p$</td>
<td>0.015</td>
<td>0.015</td>
</tr>
</tbody>
</table>

$N$: node; $M$: edge; $c$: average clustering coefficient; $L$: average path length; $k$: degree; $D$: network diameter; $p$: network density.

In a word, above by obtaining the subject categories distribution and calculating the
dynamic parameters of co-words network, we can strictly verify that generalized EQR involves both natural aspects and social aspects. The quality of the natural environment contains physical aspect, chemical aspect, biological aspect, and so on. The quality of the social environment can be divided into social-economic aspect, psychological aspect, education aspect, cultural aspect, and so on. Indeed, environmental quality, as a measure of the suitability of human survival and social or economic development, reflects the comprehensive requirements of humanity, including both natural and social demands. The calculated parameters of co-words network, especial its small world characteristics \((c=0.17\) and \(L=3.042\)), show that the connections between all aspects within the discipline is very close, holding strong correlation between the keywords and large co-occurrence chance. The pursuits for a better understanding on environmental quality must be integrated to include various aspects of EQR.

Global cooperation actions on environmental quality research (EQR)

(1) General characteristics

From the connotation of environmental quality, environmental quality is not a problem of one person or one country, but is related to the whole Earth and all humankind. We need a deep cognition on human global cooperation actions on EQR. Here we make a detailed analysis on 5154 publications that remove the articles lack of author addresses information. The results show that nearly 105 countries have conducted EQR. With countries as nodes \((N)\) and cooperation of the countries in a publication as the edge \((M)\), a cooperative network is obtained as in Figure 4, in which the size of nodes and thickness of the edges indicate the number of publications by collaboration between countries. Figure 4 visualizes the patterns of the cooperation between all countries. To quantify the characteristics of the global cooperation actions on EQR, we calculate the parameters of the network: the average clustering coefficient \(c\), the average path length \(L\), degree \(k\) and others, as listed in Table 2. The average clustering coefficient of the network has a large value (0.367), indicating the close relationship between the various countries. The average path length is 1.909, which means that on average, one country can cooperate with anyone country by no more than one intermediate country in the actions on EQR. In addition, the cooperation network has network diameter 5 and network density 0.186, indicating that global actions on EQR by cooperation between the countries is very accessible.

![Figure 7. The global cooperation network on EQR](image)
Table 2. The cooperation network of countries and calculated parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>105</td>
</tr>
<tr>
<td>M</td>
<td>1014</td>
</tr>
<tr>
<td>( c )</td>
<td>0.367</td>
</tr>
<tr>
<td>L</td>
<td>1.909</td>
</tr>
<tr>
<td>( k )</td>
<td>19.314</td>
</tr>
<tr>
<td>D</td>
<td>5</td>
</tr>
<tr>
<td>( p )</td>
<td>0.186</td>
</tr>
</tbody>
</table>

N: node; M: edge; \( c \): average clustering coefficient; L: average path length; \( k \): degree; D: network diameter; \( p \): network density.

(2) Degree distribution: characteristics of the global cooperation network on EQR
The larger the degree of a country is, the more the number of countries that cooperate with it is. From 7, the degree of USA has a largest value (80), followed by China (77) and Italy (67). The average degree of the entire network is 19.314. To reveal more underlying characteristics of the global cooperative, it is useful to study the degree distribution of the network. With \( f \) as the degrees of the countries and \( r \) as the ranks of the countries (a decreasing order), we fit the degree distribution of the network by Zipf’s equation

\[
f = c \cdot (r + m)^{-a}.
\]

The results are shown in Figure 8. \( R^2 = 0.981 \) means power-law distribution of degrees is satisfied, which implies that led by USA and China, the whole world is involved in EQR by a self-organized way. In addition, a large slope (with \( a=-3.958 \)) suggests that the number of countries with low degrees is relative large. That is, many countries have not yet sufficiently exploited their capacities in the cooperative actions on EQR.

![Figure 8. The degree distribution for all countries](image)

(3) Characteristics of cooperation between different countries and regions
In order to further understand detailed cooperative relations between different countries and regions, we here select the top 11 productive countries for our analyses.
In Figure 9, the cooperation between USA and China is most frequent, issuing 211 publications of EQR, followed by the cooperation between USA and Italy (99), USA and UK (87). USA also has wide partnerships with Canada, Spain and Brazil. Among the top 11 productive countries, in addition to China and Brazil as developing countries, the rest are all developed countries (developing countries and developed countries are identified according to the standard of 2015 United Nations). We conduct the calculations on the publications of without international cooperation (SP), international cooperation (CP) and first author (FP). The top 5 countries are still USA, China and European countries (Italy, Britain and Spain). USA accounted for 33.99% of TP, 27.74% of SP, 38.63% of CP and 22.62% of FP, followed by China (20.14% of TP, 11.57% of SP, 26.77% of CP and 12.96% of FP). Canada and Germany rank No. 6 and No. 7 in SP, higher than Brazil. Canada and France have same ranks in CP. In general, the cooperation patterns show that: a) The developed countries are led by USA and the developing countries are led by China, and both of them made much EQR. The cooperation between USA and China dominated the global patterns. b) Next to USA and China, the European countries (Italy, Britain Spain, etc) made much EQR, and there was also close cooperation between the countries. c) Brazil from South America and Canada from North America also made lots of efforts on EQR, and cooperated with many countries.

Figure 9. The cooperation of the top countries

Figure 10. Zipf’s distribution for citations

Some other factors on the global actions of EQR: citation, journal and funds

Regarding the global actions on EQR during past years, some other factors also play important roles. Here, we mainly conduct our analyses on the dynamics of citations, journals and funding agencies.

(1) Citation dynamics

In order to reveal how citation tendency and dynamics affect the global actions on EQR during past decades, we analyze the citation distribution of all 4778 articles, which is expressed by the relationship between citation times \( f \) of a publication and its rank \( r \) (the larger \( f \), the higher the rank \( r \)). We correlate the expression by a well-known generalized Zipf’s law as Eq. (2) and the results are shown in Figure 10. \( R^2=0.996 \) shows a generalized Zipf’s law or a power law distribution is well met for the citation pattern. That is, a small number of publications obtain most citations, just as the metaphor of “2/8 or 20/80 principles”. For the global actions on EQR, a
generalized Zipf’s law means the some publications can first get attention and be cited, and then they will get much more attention or be more highly cited by preferential positive feedback. Gradually a power law distribution with different citation grades will form. The preferential citation tendency and dynamics hidden in a generalized Zipf’s distribution for citation intuitively reflects the self-organized information transmission in the global actions on EQR during the past decades.

(2) The dynamics of generalized Bradford’s journal distribution

As the main exhibiting place of EQR’s publications, journals’ status, quality and distribution pattern must partly affect the global actions on EQR during past decades. In order to reveal the effect of journals on the global EQR’s actions during past decades, we analyze the journals distribution. Bradford’s law holds that the journals for scientific papers in a field have a concentration and dispersion distribution (Meadows 2002). When the journals are ranked by the number of publications in a descending order, the separation of a core area and the subsequent several areas will appear. Here we consider a generalized Bradford’s journal distribution. Taking the logarithm of the ranks of journals as horizontal axis, and the corresponding cumulative number of publications as the longitudinal axis, we have a generalized Bradford’s journal distribution for the curved portion of the figure as

\[ R(n) = an^\beta \quad (1 \leq n \leq c), \quad (3) \]

and linear portion of the figure as

\[ R(n) = k \ln\left(n \frac{N}{s}\right) (C \leq n < N). \quad (4) \]

We correlate our data using Eqs. (3-4), and the results are yielded as in Figure 11. \( R^2 = 0.992 \) and 0.998 imply that our sample well meet generalized Bradford’s journal distribution. Figure 11 shows that the journals of EQR have already included wide range of scopes. However, as an emerging interdisciplinary discipline, the participations of wider scopes of journals in dealing with EQR are still expected. In addition, since Bradford’s law expresses a centralization and decentralization journal distribution that implies a cumulative advantageous process of journal’s attraction to the submissions, we may hold the cumulative advantageous tendency hidden in a generalized Bradford’s distribution for journals reflects the self-organized information transmission in the global actions on EQR during past decades.

![Figure 11. The Bradford’s distribution](image1)

![Figure 12. The Lotka’s distribution for funds](image2)
The dynamics of funds
To some extends, fund is financial base of the global publications on EQR during past years, being an important factor for analyzing the global actions on EQR. We give the distribution dynamics of funding agencies to analyze its effect on the global EQR’s actions during past decades. We correlate the relation between the number of funding agencies \( f \) (percentage) and the number of publications \( x \) by Lotka’s expression

\[
f(x) = \frac{C}{x^{\alpha}}.
\]

In Eq. (5), \( C \) and \( \alpha \) are coefficients with different values depending on the statistical data of samples. Here, we use double-logarithmic coordinates for \( f \) and \( x \), and the results are shown in Figure 12. The slope \( (\alpha=-1.9) \) of the line for our sample basically satisfies the Lotka’s equation, which means the number of funding agency is inversely proportional to the number of publications. It is clear that Eq. (5) is another power law, which demonstrates the dynamic self-organization during the EQR from the perspective of the distribution of funding agencies. Though it is a fact that a small number of funding agencies contribute lots of efforts on EQR, the slope of the line with \( \alpha=-1.9 \) in Figure 12 means the distribution of funds is too concentrative and lots of low frequencies funding agencies have not sufficiently played their roles in the actions of EQR. Funding dynamics reflects the global actions on EQR.

Some coping strategies
The above findings can offer us some important coping strategies. Based on the connotation of EQR, since EQR lies in a small world \( (c=0.17 \text{ and } L=3.042) \), to realize a better understanding on environment quality, we should not only integrate both natural factors (water, air, soil, etc) and social factors (economy, culture, psychology, etc), but also deal with the close relationship between different aspects of EQR and strong correlation or co-occurrence between different factors. In term of the characteristics found for the global cooperative network on EQR, we should reduce the number of countries with low degrees, and make more countries be able to sufficiently plunge into the cooperative actions on EQR. Steps should be taken to improve their enthusiasms, financial supports and research capacity or efficiency. This will effectively decrease the slope of power-law distribution of degrees (currently \(|\alpha|=|-3.958|=3.958 \text{ is too large}).

As for some other factors on the global actions of EQR, we would like to stress the adjustments on the generalized Bradford’s distribution for journals and Lotka’s distribution for funding agencies. That is, EQR, as an emerging interdisciplinary discipline, should encourage wider participations of journals. The Lotka’s distribution for funding agencies with \( \alpha=-1.9 \) is too concentrative. We should encourage more low frequencies funding agencies to sufficiently play their roles in the actions of EQR, effectively decreasing the slope of Lotka’s distribution or funding agencies (currently \(|\alpha|=|-1.9|=1.9 \text{ is absolutely too large}).

Acknowledgments
The Project was sponsored by Plateau Meteorological Open Laboratory Fund
References
Overlapping Thematic Structures Extraction with Mixed-Membership Stochastic Blockmodel

Shuo Xu 1  Junwan Liu 1  Zheng Wang 2

1 {xushuo, lijunwan}@bjut.edu.cn
Beijing University of Technology, Beijing 100124, China

2 wangzheng2015@istic.ac.cn
Institute of Scientific and Technical Information of China, Beijing 100038, China

Abstract
It is increasing important to identify automatically thematic structures from massive scientific literature. The interdisciplinarity enables thematic structures without natural boundaries. In this work, the identification of thematic structures is regarded as an overlapping community detection problem from the citation-link network. Thus, the overlapping thematic structures can be detected from citation-link network with a mixed-membership stochastic blockmodel, armed with stochastic infer algorithm. Experimental results on the astro dataset indicate that it is feasible to extract overlapping thematic structures.

Conference Topic
Social Network Analysis, Mapping and Visualization

Introduction
It is increasing important to identify automatically thematic structures (topics or fields) from massive scientific literature, which is a pertinent problem of science studies in general and bibliomerics/scientometrics in particular. Last two decades witnessed significant progress in the field of the detection of thematic structures ever since van Raan (1996). Several major technologies have been developed, such as citation-link approaches (Klavans and Boyack 2011; Waltman and van Eck 2012; van Eck and Waltman 2017), lexical approaches (Leydesdorff and Welbers 2011; Yau et al. 2014), and hybrid approaches that combine citation-link and lexical techniques (Glänzel and Thijs 2011; Glänzel and Thijs 2017).

As we all know, due to intrinsic difference between these methods, they usually detect somewhat different topical structures for the same dataset. To the best of our knowledge, there is no benchmark dataset public available with known structures until now. Therefore, it is still not clear that whether the thematic structures found by these methods are indeed representative of knowledge structures in science, merely artefacts resulting from the used methods, or something in between. Though it is very difficult to answer this question, topic extraction challenge pioneered by Gläser, Glänzel and Scharnhorst (2017) as a first step in this direction tries to provide some general understanding of properties of different approaches and thematic structures they deliver.

Meanwhile, interdisciplinarity enables thematic structures without natural boundaries, which has been shown in bibliometric research (Zitt, Ramanana-Rahary & Bassecoulard 2005). So, increasing attention has been paid to detect overlapping structures from direct citation, bibliographic coupling, or co-citation networks with community detection methods in recent years (Havemann et al. 2012; Havemann, Glänzel & Heinz 2017). For more elaborate and detailed surveys on overlapping community detection we refer the readers to Xie, Kelley & Szymanski (2013).

This paper also regards the identification of thematic structures as an overlapping community detection problem from citation-link network. A mixed-membership stochastic blockmodel, armed with stochastic inference algorithm, is utilized to detect the overlapping thematic
structures. From the experimental results and comparisons with each other, it makes clear that it is feasible to discover the overlapping thematic structures.

**Mixed-Membership Stochastic Blockmodel**
Mixed-membership stochastic blockmodel (Airoldi et al. 2008; Gopalan & Blei 2013) is a variant of the stochastic blockmodel, in which each node can belong to multiple communities. Given an observed direct citation network \( \{y_{ij}\} \) (including links and non-links), hidden community/thematic structure can be discovered by estimating its conditional distribution. This model can be described by its probabilistic generative process of a network with \( K \) communities as follows. One can associate each node \( p_i \) with a multinomial distribution vector of community membership \( \theta_i \).

In order to generate an overlapping community network, this model takes into account each pair of nodes. For each pair \((i, j)\), two community indicators \( z_{i \rightarrow j} \) and \( z_{i \leftarrow j} \) are drawn from the vector \( \theta_i \) and \( \theta_j \), respectively. If these two indicators point to the same community, that is, \( z_{i \rightarrow j} = z_{i \leftarrow j} \), then it links nodes \( p_i \) and \( p_j \) with high probability \( \beta_{z_{i \rightarrow j}} \); otherwise, they are likely to be unconnected.

The above generative process defines a joint probability distribution over the node-wise community memberships \( \{\theta_i\} \), the pair-wise community indicators \( \{z_{i \rightarrow j}\} \), and the observed network \( \{y_{ij}\} \). Given an observed network \( \{y_{ij}\} \), conditional distribution (posterior) of the latent community structure \( p(\{\theta_i\}, \{z_{i \rightarrow j}\} | \{y_{ij}\}) \) gives a decomposition of the nodes into \( K \) overlapping communities. As for many interesting Bayesian models (Blei, Ng & Jordan 2003; Xu et al. 2014), however, posterior inference cannot be done exactly in this model. A variety of algorithms have been used to approximate the posterior of Bayesian models, such as variational inference (Jordan et al. 1999), Markov chain Monte Carlo (Andrieu et al. 2003), and stochastic variational inference (Hoffman et al. 2013). In this work, stochastic variational inference is used, since it can easily scale to real-world-sized citation-link networks.

**Experimental Results & Discussions**

**Dataset**
The astro dataset (Gläser, Glänzel and Scharnhorst 2017) is used in this work, which was extracted from the Web of Science (WoS) bibliographic database. It includes all publications of the document type article, letter, and proceedings paper published between 2003 and 2011 in 59 journals assigned by the Journal Citation Report to the subject Astronomy and Astrophysics. The number of publications in the dataset is 111,616. Of the 4,311,953 cited references provided in the publications of the dataset, 929,364 ones point to publications in the dataset.

Due to computational problems (Waltman & van Eck 2012; van Eck and Waltman 2017), direct citation relations ignoring directions are used here. That is to say, an undirected network is considered in this study. In the meanwhile, self-citation relations are removed (e.g., a publication citing itself). There are 659 connected components totally in the network. This work only keeps the giant component, which includes 924,750 citation relations involving \( n = 101,831 \) publications.

Before extracting the overlapping thematic structures, we want to get a first idea of the sparsity degree of the network. Let \( a_i \) denote the number of connections of node \( p_i \), and \( T = \sum_{i=1}^{n} a_i \). One can sort the \( \{a_i\} \) in the ascending order, and then let \( b_i \) represent the \( i \)-th ordered element. Gini index (GI) (Goswami, Murthy & Das 2016; Hurley & Rickard 2006) is formally calculated as follows.
\[
\text{GI} = \frac{\Delta(A)}{\Delta(A + B)} = 1 - 2\Delta(B) = 1 - 2 \left[ \sum_{i=1}^{n} \frac{b_i(n - i + 0.5)}{n} \right]
\]

Here, \(\Delta(\cdot)\) represents the area of the resulting polygon (ref. Figure 1). It is easy to see that this index is normalized and lies between 0 and 1. It measures the extent of inequality in distribution of degrees among all nodes. When GI closes to 1, it means that there are several dominant nodes in the network. For our case, GI = 0.48858 indicates that the distribution of degrees is not so skewed.

Figure 1. Gini Index of direct citations network in the *astro* dataset.

**Experimental setup**

The software *svinet*, which can be available at https://github.com/premgopalan/svinet, implements stochastic inference algorithm for the mixed-membership stochastic blockmodel. Hence, it is utilized to extract overlapping community/thematic structures from the direct citations network. The number of overlapping communities is set to \(K = 20, 30, 40\) and \(50\). The link sampling is chosen as the sampling strategy, and the prior \(\epsilon\) on network strengths is learned from the direct citations network (i.e., the option \(\text{eta-type}\) is set to \(\text{fromdata}\)).

**Overlapping thematic structures**

<table>
<thead>
<tr>
<th>No. of clusters</th>
<th>No. of pub. smallest cluster</th>
<th>No. of pub. largest cluster</th>
<th>Avg. no. of pub. per cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>6872</td>
<td>8309</td>
<td>7822.75</td>
</tr>
<tr>
<td>30</td>
<td>5272</td>
<td>6939</td>
<td>6090.23</td>
</tr>
<tr>
<td>40</td>
<td>3924</td>
<td>5531</td>
<td>4698.48</td>
</tr>
<tr>
<td>50</td>
<td>3130</td>
<td>4645</td>
<td>4072.24</td>
</tr>
</tbody>
</table>

Table 1. Statistics for the different clustering solutions

Figure 2. The distribution for the number of memberships of nodes in the *astro* dataset.
Four clustering solutions are obtained with the different values for $K$. Table 1 reports the statistics for the different clustering solutions, such as the number of clusters, the number of publications in the smallest and largest cluster, and the average number of publication per cluster. From Table 1, it is not difficult to see that, regardless of the value of $K$, the distribution of publications over clusters is quite even. The distribution of the number of memberships of nodes illustrated in Figure 2 make clear that about 40% publications only belong to one theme, and about 30% and 17% publications belong to two and three themes, respectively. This indicates that it is very common for publications to cover multiple themes. In order to get an impression of the discovered thematic structures, let’s take the solution of $K = 50$ as an example. Figure 3 visualizes the top themes with Gephi (Bastian, Heymann & Jacomy 2009) that links to the high cited publication (its degree is 1439): First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Preliminary maps and basic results. This publication bridges several themes. Each node is sized by their bridgeness (Nepusz et al. 2008), an inferred measure of their impact on multiple themes. Several publications with large bridgeness is also labelled in the Figure 3.

![Figure 3. The overlapping thematic structures in a subgraph of the astro dataset.](image)

**Solution comparisons**

Due to the fluidity of cognitive structures in science, and the multiplicity of reference frames (Gläser, Glänzel and Scharnhorst 2017), it is not easy to compare two solutions in the presence of multiple, inaccessible ground truths. But if one solution is treated as “ground truth”, then the other is seen as the solution to evaluate, and vice versa. Thus, some external measures for clustering comparison, such as pair-counting based measures, set-matching based measures, and information theoretic based measures, can be used here. Of course, the symmetric measure (switching two solutions will return the same score) should preferred. For more elaborate and detailed surveys, we refer the readers to Xu et al. (2013). Here, adjusted mutual information (AMI) (Vinh, Epps & Bailey 2010) is adopted, since it accounts for the
fact that the mutual information (MI) is generally higher for two clusterings with a larger number of clusters, regardless of whether there is actually more information shared. Table 2 shows the AMI values between our solutions, in which each publication is assigned to the theme with the highest probability. It seems that the four solutions have obvious difference, though the same clustering algorithm is used. It is very possible for AMI to evaluate the performance of overlapping thematic structures. Hence, modification of the partition coefficient (MPC) (Dave 1996) is also calculated in the last row of Table 2. The values of MPC indicate that the four solutions give similar performance.

Table 2. Adjusted mutual information and modification of the partition coefficient.

<table>
<thead>
<tr>
<th>K</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1.00</td>
<td>0.15</td>
<td>0.17</td>
<td>0.16</td>
</tr>
<tr>
<td>30</td>
<td>0.15</td>
<td>1.00</td>
<td>0.17</td>
<td>0.18</td>
</tr>
<tr>
<td>40</td>
<td>0.17</td>
<td>0.17</td>
<td>1.00</td>
<td>0.20</td>
</tr>
<tr>
<td>50</td>
<td>0.16</td>
<td>0.18</td>
<td>0.20</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>0.77</td>
<td>0.73</td>
<td>0.73</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Conclusions

Based on the astro dataset, we can conclude that the stochastic inference algorithm for the mixed-membership stochastic blockmodel can indeed scale to real-world-sized network, such as citation-link network. This approach can detect the overlapping thematic structures from massive scientific literature, and is able to identify the publications bridging several themes. What’s more, regardless of the number of themes, the distribution of publications over clusters seems to be quite even. Though the overlapping thematic structures can be discovered, the results are still not readable. We will try to label the themes with relevant phases in the next step. Additionally, in this work, only direct citation relations are utilized. We will investigate the usability of bibliographic coupling and co-citation relations in the near future.

Acknowledgments

This research received the financial support from National Science Foundation of China (ID: 71403255). Our gratitude also goes to the anonymous reviewers for their valuable comments.

References


Abstract
Bootstrap methods are increasingly accepted as one of the common approaches in constructing confidence intervals in bibliometric studies. Typical bootstrap methods assume that the statistical population is infinite. When the statistical population is finite, a correction needs to be applied in computing the estimated variance of the estimators and thus constructing confidence intervals. We investigate the effect of overlooking the finiteness assumption of the statistical population using a dataset containing all articles in Web of Science (WoS) for Delft University of Technology from 2006 until 2009. We regard the data as our finite statistical population and consider simple random samples of various sizes. Standard bootstrap methods are firstly employed in accounting for the variability of the estimates, as well as constructing the confidence intervals. The results unveil two issues, namely that the variability in the estimates does not decrease to zero as the sample size approaches the population size and that the confidence intervals are not valid. Both issues are addressed when accounting for a finite population correction in the bootstrap methods.

Conference Topic
Methods and techniques.

Introduction
The bootstrap (Efron, 1979) has become one of the most well-known statistical methods, despite the futility implied by the ‘bootstrap’ action, that is, of lifting oneself by using the bootstraps only. Nonetheless, the expressive terminology has been chosen to suggest a process done without external help. Starting from the data at hand, which is regarded as a sample from a statistical population, one is able, by resampling the data, to provide confidence intervals and means of measuring the accuracy of the sample estimates. Since the bootstrap was firstly proposed and studied, constant effort was put into improving the techniques, relaxing the assumptions and accounting for the particularities of the data at hand. This, together with the enhanced computational capabilities lead to the current popularity of the bootstrap.

Bootstrap methods are increasingly accepted as one of the common approaches to construct confidence intervals in various science fields, including citation analysis and bibliometric studies. Chen, Jen & Wu (2014), Rahman et al. (2016), Fairclough & Thelwall (2015), Williams & Bornmann (2016), Bornmann & Daniel (2007), Waltman & Costas (2013), Costas, Zahedi & Wouters (2014) are just a few recent publications that have included bootstrap methods in their analysis. Bootstrap methods are also used in accounting for the variability in the research output in the Leiden Ranking (Waltman et al., 2012).

The standard bootstrap methods, as well as various other standard statistical tools rely on the implicit assumption that the data at hand is a sample from an infinite population. “The idea of infinite hypothetical population is, I believe, implicit in all statements involving mathematical probability” (Fisher, 1925), and therefore in statistics. This is however rarely mentioned in practice by statisticians. This of course has to do with the purpose of the statistical analysis.
and the nature of the target population, as emphasized in Nane (2016) as a comment to Williams & Bornmann (2016).

Though implicit, numerous statistical populations in citation analysis are assumed to be infinite, albeit data on publications or researchers. The target population can also be regarded as a underlying stochastic process, whose realizations are observable and constitute a statistical population, see, for example Claveau (2016). The concept of super-population emerges therefore naturally.

Finite target populations arise in numerous applications, for example in survey analysis, biology, management, etc. When the target population is finite, the statistical methods need to be adapted. Hajek (1960, 1981) has initiated the work on the finite population statistics and provided important asymptotic results, as well as advanced the idea of sampling within finite populations. This study follows the discussion on the paper of Williams & Bornmann (2016). In the comment, Nane (2016) drew attention on the particularity of the data at hand, that is the finite population and the necessary correction implied by this assumption.

The purpose of this study is twofold. First of all, several techniques of performing bootstrap are considered. Results are investigated and conclusions are drawn based on the variability of the estimates and confidence intervals with respect to length and accuracy. Secondly, we will investigate the effect of not taking into account the finiteness of the population and we will, in turn, account for the finiteness of the population at hand.

The study relies on the Delft University of Technology (TU Delft) data, that is all the articles from Web of Science (WoS) for which at least one author is affiliated to TU Delft, as collected for Leiden Ranking 2011/2012. We will perform a simulation study where the data at hand will act as the statistical, finite, population and various samples, using a simple random sampling without replacement will be drawn from this population. The 6224 articles in our data have been published between 2006 and 2009 and their citations have been counted until the end of 2010. The citation performance of the articles are evaluated using two well-known indicators, MNCS and PP(top 10%). MNCS is the average of the field-normalized citation score of each publication. PP(top 10%) accounts for the proportion of publications that are in the top 10% most cited in their field. More details about the indicators can be found in Waltman et al. (2012). The population MNCS is 1.275, that is the average of the field-normalized citation scores of the TU Delft publications is 1.275, which is higher than the world average and the PP(top 10%) is 13.7%, which asserts that 13.7% of TU Delft’s publications are in the top 10% most cited publications in their field. These values will be regarded as the true parameter values and compared with estimates and confidence intervals resulting from various bootstrap methods.

The paper is organized in three subsequent sections. The following section describes the methods used for our analysis. It includes and compares standard bootstrap methods, as well as the bootstrap methods which account for the finite population. The main results of the analysis are presented in the subsequent section. Finally, the conclusions are drawn and the discussion is carried.

**Methods**

We investigate the effects of the finite population setting on the estimation of two well-known citation indicators, MNCS and PP(top 10%), as well as on the construction of confidence intervals. Standard bootstrap methods will be reviewed and compared in the following subsection. The adjustments entailed by the finite population setting will be considered and integrated in bootstrap methods.
**Bootstrap methods**

Bootstrap has become a standard statistical method for providing a measure of variation for sample estimates and computing confidence intervals. Since it was first advanced by Efron (1979), numerous bootstrap methods have been proposed.

Given an independent and identically distributed sample \( s = \{x_1, x_2, \ldots, x_n\} \) from an unknown distribution \( F \), we are interested in estimating the parameter \( \theta \) of the distribution \( F \), typically the mean. The estimate of \( \theta \) is denoted by \( \hat{\theta} \), the sample mean estimate of the mean of the distribution. \( B \) samples, denoted \( s_1^*, s_2^*, \ldots, s_B^* \), are drawn, with replacement from the initial sample \( s \). These samples represent the bootstrap samples and the size of each bootstrap sample is equal to \( n \). A bootstrap parameter \( \theta_i^* \) is computed for every bootstrap sample \( s_i^* \), for \( i = 1, \ldots, n \).

The variance of the estimator \( \hat{\theta} \) is estimated by the bootstrap variance estimator. In turn, the bootstrap variance estimator has been shown to converge asymptotically to the sample variance of the bootstrap estimator.

This bootstrap methods is also referred to as the nonparametric bootstrap, since no parametric distribution is assumed for \( F \). Assuming a certain parametric family for \( F \) and drawing the bootstrap samples from a parametric distribution where the parameters have been estimated from the sample \( s \) are often referred to as the parametric bootstrap. We will not include the parametric bootstrap in our analysis.

The standard approach for computing confidence intervals is based on the central limit theorem, and uses the sample variance bootstrap estimator and quantiles of the standard normal distribution. These confidence intervals will be called asymptotic confidence intervals. Confidence intervals can also be constructed based on the percentiles of the bootstrap distribution, suggestively named percentile confidence intervals. The percentiles confidence intervals can be quite inaccurate in the case of skewed distributions. A bias corrected version (BCa) of the bootstrap has been proposed to compute confidence intervals (Efron, 1984). DiCiccio & Efron (1996), for example, provide a detailed theoretical analysis of the accuracy of these bootstrap confidence intervals.

Though implicit, standard bootstrap methods assume that the statistical population from which the data at hand was sampled is infinite. In the following section, we will see that, in practice, bootstrap methods fail to provide adequate results when this assumption is violated.

**Finite population correction**

The finite population correction assumes that the population size is known. Let \( N \) be the population size and let \( n \) be the sample size. Then the sampling fraction is defined as \( f = n/N \) and the finite population correction is given by \( 1 - f \), see, for example, Davidson and Hinkley (1997). If the variance of the bootstrap estimator is denoted by \( V^* \), then the finite population correction of the variance estimator is \( (1 - f)V^* \), from where it becomes obvious that it converges to zero, as \( n \to N \). A straightforward bias-adjusted variance estimator is given by

\[
V^{**} = \frac{N - n}{N - 1} V^*.
\]

When estimating means, the standard error of the mean for finite populations is given
\[ \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \]

The second term of the product is often referred to as the finite population correction factor (fpc) and is also used in deriving the standard error of the proportion for finite populations. Finally, it needs to be emphasized that the finite population correction depends on the sampling design. As mentioned beforehand, the employed sampling scheme here is simple random sampling without replacement. If a different sampling design is chosen or has been already employed when collecting the data, then the finite population correction needs to be adjusted accordingly.

The question is of course when does the finite population make a significant difference. Davidson and Hinkley (1997) advise to use the finite population correction when \( f \geq 0.1 \). That is to say, if \( f < 0.1 \), then \( n \) is relatively small compared to \( N \) and the correction factor can be ignored.

**Bootstrap methods for finite populations**

Several methods have been proposed in the literature for the bootstrap while taking into account the finiteness of the population. Mashreghi, Haziza & Léger (2016) have recently provided a survey of bootstrap methods for finite populations in the context of survey data and for various sampling techniques. We will investigate two bootstrap methods, namely the pseudo-population bootstrap and the direct bootstrap for simple random sampling without replacement.

The pseudo-population bootstrap method creates, by resampling, a pseudo-population with the same size as the actual finite population. Bootstrap samples, equally sized as the original sample are subsequently drawn from the resulting pseudo-population. By employing the same sampling design in drawing the bootstrap samples from the pseudo-population as for drawing the initial sample, the method ensures that the bootstrap variance estimator encompasses the finite population correction factors. The pseudo-population method has been first proposed by Gross (1980) and several adaptations have been developed ever since. We will employ in our analysis the method proposed by Booth, Butler & Hall (1994).

For comparison reasons, we have also employed the direct bootstrap method for finite populations. Direct bootstrap methods are grouped in a category that does not employ the creation of the pseudo-population but mimic the idea and adjust the original proposal of Efron (1979), where bootstrap samples are directly drawn from the data at hand. The adjustments are made “so that the bootstrap variability reflects the sampling variability of the original sample design” (Mashreghi, Haziza & Léger, 2016) while accounting for the finite population setting. In the direct method proposed by Sitter (1992), resamples of smaller size are taken from the initial sample without replacement for a number of times that depends on the size of the original sample and the resample and the finite population correction factor. The bootstrap sample is constructed by concatenating all these resamples. The motivation and technical details, along with two other direct methods are extensively studied in Mashreghi, Haziza & Léger (2016). A comparison between the direct bootstrap methods and the standard bootstrap proposed by Efron (1979) shows how Efron’s bootstrap method, in the absence of a finite population correction factor, leads to an overestimate of the variance.

Confidence intervals can be constructed using the same techniques as when using the standard bootstrap methods. Parametric confidence intervals will be constructed, as well as bootstrap-t confidence intervals. The bootstrap-t confidence intervals use quantiles of the bootstrap
distribution \((\hat{\theta} - \theta)/\sqrt{\hat{V}}\), where \(\hat{V}\) is an estimate of the variance of the estimator \(\hat{\theta}\) instead of quantiles of the standard normal distribution. Moreover, we will employ standard percentile confidence intervals that make use of the bootstrap distribution of \(\hat{\theta}^*\). As noted by Mashreghi, Haziza and Léger (2016), it is worth mentioning that, in practice, the percentile confidence intervals over-cover the true parameter \(\theta\). This is generated by the overall high dispersion of the bootstrap estimators \(\hat{\theta}^*\) compared to the dispersion of \(\hat{\theta}^* - \theta^*\) that is used in computing the bootstrap-t confidence intervals.

**Finite population in statistical software packages**

The assumption of finite population is usually contained within the survey analysis of the statistical software programs. The finite population correction is included in various statistical software programs as Stata, SAS and R. In Stata, it is included as an option in the survey design functions. SAS also has the finite population correction factor contained within the survey function PROC SURVEYMEANS. R has various packages that account for finite population correction, including survey and PracTools. In Stata, the variance estimation and confidence intervals for finite population are obtained using bootstrap procedure proposed by McCarthy & Snowden (1985) and Rao & Wu (1988) in the class of direct bootstrap methods. The differences between the two methods and the method of Sitter (1992) employed in our analysis are quite technical and can be found in the survey of Mashreghi, Haziza & Léger (2016). To the best of our knowledge there are no bootstrap methods for finite populations implemented in SAS. In R, the survey package also the direct bootstrap method of Rao & Wu (1988). To the best of our knowledge, the pseudo-population bootstrap and the direct bootstrap method that we used in our analysis have not been implemented yet in a package in R or elsewhere.

**Results**

This section includes the analysis of the standard bootstrap method for various sample sizes, as well as two bootstrap methods that allow for the finite population assumption.

**Standard bootstrap methods**

To study the sample estimates and to compare the different 95% confidence intervals, a sample of size \(n = 100\) has been drawn from our statistical population. The table below shows the results for the two indicators, MNCS and PP(top 10%).

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Variance of Estimator</th>
<th>Asymptotic CI</th>
<th>Percentile CI</th>
<th>BCa CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNCS</td>
<td>1.22</td>
<td>0.21</td>
<td>(0.88;1.71)</td>
<td>(0.91;1.74)</td>
</tr>
<tr>
<td>PP(top 10%)</td>
<td>12.58%</td>
<td>0.03</td>
<td>(6.44%;18.89%)</td>
<td>(6.69%;19.12%)</td>
</tr>
</tbody>
</table>

The sample MNCS is 1.22 and the estimate for PP(top 10%) is 12.58%, which are relatively different from the true values of MNCS is 1.275 and PP(top 10%) is 13.65%. The variance of the two estimators is estimated using the bootstrap method. As expected, the variance for the MNCS is significantly higher than the variance for the PP(top 10%).
difference reflects the robustness of the PP(top 10%) as compared to the MNCS. The three confidence intervals are rather wide, showing the high uncertainty around the estimates. Finally, it is notable that all confidence intervals contain the true values. Obviously these results depend on the particular initial sample, so an analysis of the bootstrap has been performed by repeating the bootstrap procedure for a large number of times. Specifically, for 1000 times, a sample of size 100 has been drawn. The MNCS and PP(top 10%) estimate has been computed, along with the variance of the estimators and the bootstrap confidence intervals for the 1000 samples. The results are depicted in Table 2.

Table 2. Average length and coverage of the true value for asymptotic, percentile and BCa confidence intervals (CI) for MNCS and PP(top 10%) based on a sample n=100 and B=1000 bootstrap samples.

<table>
<thead>
<tr>
<th>Average of Variance of Estimator</th>
<th>Type CI</th>
<th>Average length</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNCS</td>
<td>Asymptotic</td>
<td>0.71</td>
<td>89.8%</td>
</tr>
<tr>
<td></td>
<td>Percentile</td>
<td>0.76</td>
<td>90.7%</td>
</tr>
<tr>
<td></td>
<td>BCa</td>
<td>0.81</td>
<td>91.9%</td>
</tr>
<tr>
<td>PP(top 10%)</td>
<td>Asymptotic</td>
<td>12.89</td>
<td>91.5%</td>
</tr>
<tr>
<td></td>
<td>Percentile</td>
<td>12.9</td>
<td>93.7%</td>
</tr>
<tr>
<td></td>
<td>BCa</td>
<td>13.27</td>
<td>95.8%</td>
</tr>
</tbody>
</table>

The average of the estimated variance of the estimator has been obtained by averaging the estimated variances of the two estimators over the bootstrap samples. We note small differences from the estimated variance of the estimator using one sample in Table 1. The average length provides the average over the length of the confidence intervals obtained from applying the three bootstrap methods. Furthermore, we computed the coverage probability of the confidence intervals. That is, we checked whether the true MNCS and PP(top 10%) values were contained in each resulting confidence interval. We concluded that, for example, the true MNCS values was contained in 907 out of 1000 percentile confidence intervals. Similarly, 958 times was PP(top 10%) true value contained in the BCa confidence intervals. Since we computed 95% confidence intervals, we expect a 95% coverage of the confidence intervals. The results in Table 2 show that the 95% confidence intervals are mostly under-covering the true values. This effect is highly likely to be due to the small sample size. We will hence investigate the coverage probability, as well as the average length and average of the estimated variance of the estimator for larger samples, namely for \( n = 1000 \).

Table 3. Average length and coverage of the true value for asymptotic, percentile and BCa confidence intervals (CI) for MNCS and PP(top 10%) based on a sample n=1000 and B=1000 bootstrap samples.

<table>
<thead>
<tr>
<th>Average of Variance of Estimator</th>
<th>Type CI</th>
<th>Average length</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNCS</td>
<td>Asymptotic</td>
<td>0.27</td>
<td>96.4%</td>
</tr>
<tr>
<td></td>
<td>Percentile</td>
<td>0.27</td>
<td>96.1%</td>
</tr>
<tr>
<td></td>
<td>BCa</td>
<td>0.28</td>
<td>96.8%</td>
</tr>
<tr>
<td>PP(top 10%)</td>
<td>0.01</td>
<td>Asymptotic</td>
<td>4.13</td>
</tr>
<tr>
<td>------------</td>
<td>------</td>
<td>------------</td>
<td>------</td>
</tr>
<tr>
<td>Percentile</td>
<td>4.10</td>
<td>98.2%</td>
<td></td>
</tr>
<tr>
<td>BCa</td>
<td>4.13</td>
<td>99.1%</td>
<td></td>
</tr>
</tbody>
</table>

The results in Table 3 show a rapid decrease of the variance of estimators and average length of the confidence intervals as the sample size increases. The average of the estimated variance of the estimators has decreased to less than a half. The average length of the confidence intervals has also decreased considerably. Finally, we notice an increase in the confidence intervals coverage. The high coverage values, which might be interpreted as a very good performance, show that the confidence intervals are conservative and over-cover the true parameter.

It also raises a question on the validity if the confidence interval. A confidence interval is said to be valid if the coverage of the interval converges to the true coverage (of 95% in our case) for a very large number of repetitions of the procedure. When repeating the procedure for a number of times, the coverage probabilities have slightly changed.

Furthermore, when repeating the bootstrap methods for increasing sample sizes, the variance of the estimators and the length of the confidence intervals decreases. Since the asymptotic confidence intervals are derived using the variance of the estimator, we will investigate the influence of the sample size on the length of the asymptotic confidence intervals. Figure 1 below depicts the length of the confidence intervals for sample sizes until 4000 observations and for a sample size equal the statistical population of 6224 observations.

![Figure 1. Length of asymptotic confidence intervals for MNCS for different sample sizes using the standard bootstrap method (left) and length of 100 MNCS confidence intervals using a sample size equal to the population size (6224).](image)

The two plots show that, although decreasing, the variance of the MNCS estimator and hence the confidence intervals do not decrease to zero, even when the sample size equals the size of the statistical population, as depicted in the right plot. When the sample is, in fact, the statistical population, the estimate is as a matter of fact the true parameter and has zero variance. Nonetheless, this is not reflected by the right plot of Figure 1. The reason why the variance and hence the length of the confidence intervals is not decreasing to zero is because the sample is assumed to be drawn from an infinite population. A finite
population correction therefore needs to be applied. The bootstrap methods that account for finite populations are employed in the following subsection.

**Bootstrap methods for finite population**

We will use the pseudo-population and the direct method described in the previous section for constructing asymptotic, bootstrap-t and percentile confidence intervals for the MNCS and PP(top 10%) estimators. The results for a simple random sample of size 500 are included in Table 4.

<table>
<thead>
<tr>
<th>Table</th>
<th>Bootstrap Type CI</th>
<th>Average CI</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNCS</td>
<td>PPB Asymptotic</td>
<td>0.37</td>
<td>93.1%</td>
</tr>
<tr>
<td></td>
<td>Bootstrap-t</td>
<td>0.36</td>
<td>92.4%</td>
</tr>
<tr>
<td></td>
<td>Percentile</td>
<td>0.36</td>
<td>94.9%</td>
</tr>
<tr>
<td></td>
<td>DB Asymptotic</td>
<td>0.37</td>
<td>95.2%</td>
</tr>
<tr>
<td></td>
<td>Bootstrap-t</td>
<td>0.37</td>
<td>92.3%</td>
</tr>
<tr>
<td></td>
<td>Percentile</td>
<td>0.37</td>
<td>95.1%</td>
</tr>
<tr>
<td>PP(top 10%)</td>
<td>PPB Asymptotic</td>
<td>5.58</td>
<td>95.6%</td>
</tr>
<tr>
<td></td>
<td>Bootstrap-t</td>
<td>5.85</td>
<td>95.2%</td>
</tr>
<tr>
<td>PP(top 10%)</td>
<td>DB Percentile</td>
<td>5.57</td>
<td>95.8%</td>
</tr>
<tr>
<td></td>
<td>Asymptotic</td>
<td>5.67</td>
<td>94.4%</td>
</tr>
<tr>
<td></td>
<td>Bootstrap-t</td>
<td>5.68</td>
<td>94.1%</td>
</tr>
<tr>
<td></td>
<td>Percentile</td>
<td>5.67</td>
<td>95.1%</td>
</tr>
</tbody>
</table>

We observe similar results in terms of the average length of the intervals for the two bootstrap methods. The coverage of the confidence intervals suggest, in general, a better coverage for the PP(top 10%) than for MNCS. When using the pseudo-population bootstrap, the MNCS seems to be under-covered by the confidence intervals, where the highest coverage probability is 94.9% for the percentile confidence intervals. Using the direct method produces confidence intervals that cover the true MNCS value close to the nominal coverage. For PP(top 10%), the coverage probabilities for the pseudo-population confidence intervals are closer to the nominal value. The closest coverage probabilities to the nominal value, for both MNCS and PP(top 10%), and for both bootstrap methods are obtained by the percentile confidence intervals.

<table>
<thead>
<tr>
<th>Table</th>
<th>Bootstrap Type CI</th>
<th>Average CI</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNCS</td>
<td>PPB Asymptotic</td>
<td>0.24</td>
<td>93.1%</td>
</tr>
<tr>
<td></td>
<td>Bootstrap-t</td>
<td>0.24</td>
<td>92.4%</td>
</tr>
<tr>
<td></td>
<td>Percentile</td>
<td>0.24</td>
<td>94.9%</td>
</tr>
<tr>
<td></td>
<td>DB Asymptotic</td>
<td>0.25</td>
<td>95.2%</td>
</tr>
<tr>
<td></td>
<td>Bootstrap-t</td>
<td>0.24</td>
<td>92.3%</td>
</tr>
<tr>
<td></td>
<td>Percentile</td>
<td>0.25</td>
<td>95.1%</td>
</tr>
</tbody>
</table>
Table 5 shows the analysis for a simple random sample of 1000 observations. The average length is decreasing when comparing with Table 4. Nonetheless, the coverage probability fluctuates around the nominal coverage probability of 95%, sustaining the validity of the confidence intervals produced. Once more, the percentile confidence intervals are the closest to the true coverage probability.

To investigate the change in the length of the confidence intervals with respect to the sample size, we performed simulations for sample sizes varying from 100 to the entire statistical population.

Figure 2. Length of asymptotic and percentile confidence intervals (CI), for n=100, 500, 1000, 2000, 4000 and 6224 using the direct bootstrap method (left), pseudo-population bootstrap (middle) and the percentile confidence intervals produced with boot.ci function in R (right).

Figure 2 above shows the length of the asymptotic and percentile confidence intervals using samples with size varying from 100 (leftmost) to 6224, the size of the statistical population. For each sample size, the direct method of Sitter (1992), the pseudo-population of Booth, Butler & Hall (1994) and the standard bootstrap method from Davidson & Hinkley (1997) implemented in the statistical R package boot are used to compute confidence intervals. As expected, the length of the confidence intervals decreases as the sample size increases. It is notable however that the standard bootstrap method used by the boot.ci function in R does not decrease to zero.

Conclusions and discussion

The purpose of this study is to draw attention and provide a detailed analysis for the finite populations when bootstrap techniques are employed. The confidence intervals constructed using standard bootstrap techniques have been shown to over-cover the true parameter as the sample size increases. The over-coverage appears for sample sizes of less than 16% of the statistical population. The over-coverage poses serious questions on the validity of the confidence intervals, which need to have a coverage around the
nominal value of 95%. The main reason for this phenomenon is that the variability in the estimator does not take into account the finiteness of the population. The analysis shows that the estimated variance of MNCS and PP(top 10%) estimates decreases as the sample size increases. Despite small, the estimates of the variance of MNCS and PP(top 10%) influence significantly the coverage of the confidence intervals.

Finite populations depend on the choice of target population and can definitely occur in citation analysis. An example at hand is the use of the bootstrap methods to compute the stability intervals in the Leiden Ranking (Waltman et al., 2012). All publications of universities in WoS are regarded as the (finite) target population from which samples of size 1000 are drawn, with replacement, in order to compute the stability intervals. It is notable that the widest stability intervals are encountered for universities with the smallest output, of around 1000 publications. On the other hand, for these universities, the samples are the largest with respect to the population size, which should translate into smaller variation of the indicators and hence smaller stability intervals. A finite population correction for the variability of the indicators, as well as bootstrap methods for finite population would account for the relationship between the sample size and the population size.

The results suggest that the finite population correction needs to be incorporated in any proper analysis. Nonetheless, the scarce and limited availability of bootstrap methods for finite populations make the process difficult. Despite the large application of these techniques, they are contained in packages and functions devoted to survey analysis. Though far from exhaustive, the bootstrap methods presented in this study are theoretically accurate, as presented by Mashreghi, Haziza & Léger (2016). We need to emphasize though that a large number of variations of the bootstrap are available, depending on the focus of the analysis and the data at hand.

First of all, the sampling procedure itself can lead, in the case of finite populations to different bootstrap methods. In our study, we used simple random sampling without replacement. Other methods are available, including mirror-match bootstrap and superpopulation bootstrap. The details of these methods can be found in Sitter (1992) and Davidson & Hinkley (1996). All in all to show that finite population sampling is not a trivial matter and the effect of the sampling design should be accounted for. In practice, researchers can opt for stratified sampling, unequal sampling, etc. The study of the effect of the sampling design on the variability of the bibliometric indicators would be of interest.

An important assumption of the methods presented is that the data at hand are independent and identically distributed. It has been frequently shown that the independence assumption is unrealistic and one can argue that, in this analysis, the citation counts of the publications at TU Delft are not independent. Bootstrap methods are available for dependent, though identically distributed data (Gonçalves & Politis, 2011) and it would be very interesting to investigate the existing dependencies and also the effect of those dependencies on the estimates and confidence intervals.

Last, but definitely not the least, we reinforce the concepts of statistical sample and statistical population. A very important question evidently aims for the target of the analysis and therefore target population. As emphasized in Nane (2016), the target population triggers the analysis and therefore the validates the use of finite population correction and bootstrap methods of computing the confidence intervals.

Acknowledgments
The authors are grateful to CWTS, Leiden University, for providing the data for this study.
References


Further examples of dominance structure measurement

Yuxian Liu and Ronald Rousseau

1 yxliu@tongji.edu.cn
Tongji University Library, Tongji University, Siping Street 1239,
Yangpu District, Shanghai 200092, China

2 ronald.rousseau@uantwerpen.be
University of Antwerp, Faculty of Social Sciences, Antwerpen, Belgium

Abstract
In this article we continue our study of power structures, providing further examples of dominance structure in a directed acyclic network. We calculate the D-measure, a measure expressing the degree of dominance in a network, when nodes are added to an existing simple network. This was done for “moving from a hand to a line”, and for a citation network consisting of the ego (the original article) with n references and m received citations. When n=m the network is symmetric and the D-measure is 0.5; when the article did not yet receive a citation the D-measure is n/(n+1). When the article has no references then the D-measure is 1/(m+1).

Keywords: dominance structure; power; digraph; citation graph

Introduction
The universe is often viewed in a dual fashion: where there is high there is low; where there is left there is right; where there is yin there is yang; where there is dominance, there is subdominance. Yet, each of these two opposing principles are usually not present in equal amounts. We will explain that these notions can be expressed in mathematical terms using zero-sum arrays (Liu, Rousseau & Egghe, 2017). More specifically, these zero-sum arrays will be used to study dominance and subdominance.

It goes without saying that network theory is an essential part of informetrics; the best known networks in the field being direct citation networks, co-citation networks and other co-occurrence networks such as co-word or co-assignation networks. These networks reveal key aspects of the structure of science and as such studies of their characteristics belong to the core of informetrics. Among the many characteristics of networks we consider here dominance, defined in detail further on. Scientific disciplines benefit if networks making part of the discipline are studied from many angles, dominance and sub-dominance being one of the many aspects. As science is a formal as well as an informal structure also on this higher level, it includes dominance structures, which may change when top scientists change affiliations.

Power structures are indeed ubiquitous. Blogs and information systems are modern forms of social power (Wei, 2009; Liu et al., 2013; Lu et al., 2015; Nord et al., 2016), while universities and systems of education too can be described as power structures (Clark, 1987; Van de Graaff et al., 1978). Power structures among academics may influence scholarly impact in research (Truex et al., 2009). Empowerment of women, in academia, in industry, in politics and in daily life is another topic that has lately received quite some attention (Gutierrez et al., 2000; Pacardo-Mercado, 2013; Oreglia & Srinivasan, 2016). Yet, the majority of these studies are qualitative in nature. Hence, a quantitative framework for measuring power is necessary. This is provided in (Liu et al., 2017). In this earlier work we discussed local and global dominance as special network structures. Next we focused on dynamic aspects of dominance structures (Liu & Rousseau, 2017). In this investigation we continue our study of dynamic aspects of dominance structures adding nodes in a network one
at the time. Studying dynamic aspects of networks is essential for potential applications. We recall that in our previous studies and also here the focus goes to structures, not elements or single nodes.

**Zero-sum arrays**

A zero-sum array is defined as follows: If $X = (x_1, \ldots, x_N)$ is a real-valued array, i.e. an $N$-tuple, such that $\sum_{i=1}^{N} x_i = 0$, then $X$ is called a zero-sum array. The set of all zero-sum arrays is denoted as $\mathbf{Z}$; its subset of arrays of length $N$, namely $\mathbf{Z}_N$, is denoted as $\mathbf{Z}_N$. The array $Y_1 = (4,3,-7)$ is an example of a zero-sum array and so is $Y_2 = (4,3,0,-7)$. In accordance with the theory developed in (Liu et al., 2017) all arrays studied in this article must be ranked from highest to lowest value.

Constructing a pseudo-Lorenz curve for zero-sum arrays.

Given a zero-sum array $X$, we set

$$I_+(X) = \{i \in \{1,\ldots,N\} \text{ such that } x_i > 0\}$$
$$I_0(X) = \{i \in \{1,\ldots,N\} \text{ such that } x_i = 0\} \text{ and}$$
$$I_-(X) = \{i \in \{1,\ldots,N\} \text{ such that } x_i < 0\}.$$

As in (Liu et al., 2017) we assume that $X$ is not the trivial zero-array. This implies that $I_+(X)$ and $I_0(X)$ are always non-empty. This requirement also implies that $N > 1$. When it is clear about which array we are talking or when it does not matter we simply write $I_+, I_0$ or $I$. We note that $\sum_{i \in I} x_i = -\sum_{i \in I} x_i$. Next we put $\Sigma_+ = \sum_{i \in I_+} x_i$ and $\forall i = 1,\ldots,N : a_i = \frac{x_i}{\Sigma_+}$. With each zero-sum array $X$, we associate a corresponding $A$-array, denoted $AX$, and equal to $AX = (a_1, \ldots, a_N)$. We will further need the array $Q_X$, with $(Q_X)_i = q_i = \sum_{k=1}^{i} |a_k|$. Clearly $q_N = 2$.

Liu et al. (2017) developed a new curve, referred to as the D-curve (D for dominance), which does not decrease when the corresponding D-array has negative values, but, instead, rises further to the value two. Let us recall this construction of a D-curve.

A D-curve of a zero-sum array $X$ is defined as the polygonal line connecting the points:

$$(0,0) \rightarrow \left( \frac{1}{N}, a_i \right) \rightarrow \ldots \rightarrow \left( \frac{i}{N}, \sum_{j=1}^{i} a_j \right) \rightarrow \ldots \rightarrow \left( \frac{|I_+|}{N}, 1 \right) \rightarrow \left( \frac{N - |I_-|}{N}, 1 \right) \rightarrow \ldots \rightarrow \left( \frac{k}{N}, \sum_{j=1}^{k} |a_j| \right) \rightarrow \ldots \rightarrow (1,2)$$
where $i \in I_+, k \in I_-$. We see that a D-curve is partly concave and partly convex, see Fig.1., representing the D-curve of the array $(4,2,0,0,0,-1,-1,-4)$.

Moreover, this curve can be described as a function, denoted as $D_X(t)$: for $t \in [0,1]$, we have:

$$D_X(t) = \begin{cases} 
Na_i t, & t \in [0, \frac{1}{N}] \\
\sum_{k=1}^{i} a_k + Na_{i+1} \left( t - \frac{i}{N} \right), & t \in \left[ \frac{i}{N}, \frac{i+1}{N} \right], \quad i = 1, \ldots, N-|I|-1 \\
\sum_{k=1}^{i} |a_k| + Na_{i+1} \left( t - \frac{i}{N} \right), & t \in \left[ \frac{i}{N}, \frac{i+1}{N} \right], \quad i = N-|I|, \ldots, N-1 
\end{cases}$$

![Figure 1. An example of a D-curve](image)

Similar to Lorenz curves, also D-curves introduce a partial order, called the dominance relation.

**Definition:** the dominance relation $\preceq_D$ in $\mathbb{Z}$

Let $X$ and $Y$ be zero-sum arrays, not necessarily of the same length, then we say that $X$ is D-smaller than $Y$ (or $Y$ is D-larger than $X$), denoted as $X \preceq_D Y$ (or $Y \succeq_D X$) if, for each $t$, $D_X(t)$...
The array \( X \) is strictly \( D \)-smaller than the array \( Y \), denoted as \( X \prec_D Y \), if, for each \( t \) \( D_X(t) \leq D_Y(t) \) and there is at least one point \( t_0 \), hence infinitely many, where \( D_X(t_0) < D_Y(t_0) \). The relation \( \leq_D \) determines a partial order in the set of all equivalence classes (all arrays with the same \( D \)-curves) of zero-sum arrays. Formally we write:

\[
X \leq_D Y \quad \text{if and only if} \quad \forall t \in [0,1]: D_X(t) \leq D_Y(t)
\]

As the dominance relation \( \leq_D \) is only a partial order, some arrays cannot be compared: they are said to be intrinsically incomparable.

The area between the \( D \)-curve and the line \( y = x \) is denoted as \( AR_D(X) \). For any zero-sum array this measure takes values on the interval \( ]0,1[ \). We will refer to this measure as the \( D \)-measure. The \( D \)-measure is denoted \( AR_D \) and is calculated as:

\[
AR_D(X) = \frac{1}{N} \left( \sum_{i=1}^{N} q_i \right) - \frac{N+2}{2N}
\]

(1)

where the \( q \)-values are the components of the array \( Q_X \): \( (Q_X)_j = q_j = \sum_{k=1}^{j} |a_{jk}| \). For further information on \( D \)-curves we refer the reader to (Liu et al., 2017; Liu & Rousseau, 2017).

It was shown in (Liu et al., 2017) that this zero-sum theory can be applied in directed, acyclic networks. We briefly recall its main features.

The number \( \alpha_j^+ \) of edges in the digraph \( G \) having node \( j \) as their initial node is called the out-degree of node \( j \). Similarly, the number \( \alpha_j^- \) of edges in \( G \) having node \( j \) as their terminal node is called the in-degree of node \( j \). We put \( \alpha_j = \alpha_j^+ - \alpha_j^- \). This parameter characterizes the flow through node \( j \). Since every edge is outgoing from a node and terminating at another, the number of edges in \( G \), denoted as \( \varepsilon \), is related to the degrees of its nodes by the following equation:

\[
\varepsilon = \sum_j \alpha_j^+ = \sum_j \alpha_j^- \quad \text{or} \quad \sum_j \alpha_j = 0
\]

where the summation is over all nodes of graph \( G \). Hence the sequence \( (\alpha_j) \) forms a zero-sum array derived from the network and we can apply the \( D \)-theory explained above. In this contribution we will, however, not study local flows but focus on global flows and the corresponding global dominance theory, explained below.

In global dominance theory, GDT in short, arrays of the form \( \Sigma = (\sigma_1, \sigma_2, \ldots, \sigma_N) \), are used, defined as follows:

\[
\sigma_i^+ = \text{the sum of the lengths of the chains that start in node } i
\]

\[
\sigma_i^- = \text{the sum of the lengths of all chains that end in node } i
\]

\[
\sigma_i = \sigma_i^+ - \sigma_i^-
\]
From these definitions we see that also $\Sigma$ leads to a zero-sum array and hence also here we can apply the zero-sum theory developed in (Liu et al., 2017). This zero-sum array will be referred to as a global flow array and consists of global flow numbers.

**Modelling: the need for examples and case studies**

When a new measure is proposed one usually derives theoretical properties and explains the possible benefits of using such a measure. This has been done in (Liu et al., 2017). Studying dynamic aspects of networks and their corresponding dominance measures is essential for potential applications in fields such as business management, politics and social interactions. We provided examples in politics (Sales, 1991), universities and systems of education (Clark, 1987; Van de Graaff et al., 1978). We further mentioned the important case of changing power structures when two companies merge.

**The global D-measure: Two case studies**

We next determine and study the global D-measure for two interesting dynamic structures.

**Case I. From a hand to a line**

We consider a hand with $N-1$ fingers, i.e. with $N$ nodes in total. This hand evolves to a line as shown in Figure 2. We will calculate values of the D-measure for the cases a), b) and c).

![Figure 2. From a hand to a line, illustrated for $N = 6$.](image)

Case studies are considered with $N$ nodes in total; the one on top has global flow value $f_1$ and then further numbered downwards.

Case (a). This “hand” leads to a maximum D-graph. It is shown in (Liu et al., 2017) that the corresponding D-measure is equal to $(N-1)/N$.

Case (b). With at least 3 nodes.
Top: $f_1 = 2(N-2) + 1$

Middle (split): $f_2 = (N-2) - 1 = N-3$

Rest: $f_3$ till $f_N$ (N-2 nodes in total): -3 each

The positive part is: $3N-6$ (if $N>2$)

$|A|\text{-array: } \begin{pmatrix} \frac{2N-3}{3N-6} \frac{N-3}{3N-6} \frac{3}{3N-6} \end{pmatrix}_{(N-2) \text{ times}}$

$Q\text{-array } = \begin{pmatrix} \frac{2N-3}{3N-6}, \frac{1}{3N-6}, \frac{(3N-6) + 3i}{3N-6} \end{pmatrix}_{i = 1,\ldots,N-2}$

$AR_{D}(X) = \begin{pmatrix} 1 & 1 \\ \frac{3N-6}{N} & \frac{N+2}{2N} \\ \frac{N}{3N-6} & \frac{9N^2 - 23N + 12}{2N} & \frac{N+2}{2N} \\ \frac{9N^2 - 23N + 12}{2N} & \frac{(N+2)(3N-6)}{2N(3N-6)} \\ \frac{6N^2 - 23N + 24}{6N(N-2)} \end{pmatrix}$

Case (c). With at least 4 nodes.

Top: $f_1 = 3(N-3) + 2 + 1 = 3N - 6$

Second: $f_2 = 2(N-3) + 1 - 1 = 2N-6$

Node at split: $f_3 = -3 + (N-3) = N-6$

Rest: $f_4$ till $f_N$ (N-3 in total): -6 each

The positive part is: $6N-18$ (if $N>5$; other cases are not included here)

$|A|\text{-array: } \begin{pmatrix} \frac{3N-6}{6N-18} \frac{2N-6}{6N-18} \frac{N-6}{6N-18} \frac{6}{6N-18} \end{pmatrix}_{(N-3) \text{ times}}$
Q-array = \left( \frac{3N-6}{6N-18}, \frac{5N-12}{6N-18}, \frac{1}{6N-18}, (\frac{6N-18}{6N-18})^i \right), i = 1, ..., N-3

\text{AR}_D(X) =
\begin{align*}
\frac{1}{N} \frac{1}{6N-18} \left( \frac{3N-6}{6N-18} + (5N-12) + (6N-18)(N-2) + \frac{6(N-3)(N-2)}{2} \right) - \frac{N+2}{2N} \\
= \frac{1}{N} \frac{1}{6N-18} \left( 3N-6 + 5N-12 + 6N^2 - 18N - 12N + 36 + 3N^2 - 15N + 18 \right) - \frac{N+2}{2N} \\
= \frac{1}{N} \frac{1}{6(N-3)} \left( 9N^2 - 37N + 36 - (N+2)(3N-9) \right) = \frac{9N^2 - 37N + 36 - 3N^2 + 3N + 18}{6N(N-3)} \\
= \frac{3N^2 - 17N + 27}{3N(N-3)}
\end{align*}

\text{Case (d). With at least 5 nodes.}

\text{Top: } f_1 = 4(N-4) + (3 + 2 + 1) = 4N - 10

\text{Second: } f_2 = 3(N-4) + (2+1) - 1 = 3N-10

\text{Third: } f_3 = 2(N-4) + 1 - (1+2) = 2N - 10

\text{Node at split: } f_4 = -(3+2+1) + (N-4) = N-10

\text{Rest: } f_5 \text{ till } f_N \text{ (N-4 in total): } -(4+3+2+1) = -10 \text{ each}

\text{The positive part is } 10N-40 \text{ (if } N>9; \text{ other cases are not included here)}

[A]-array:
\begin{bmatrix}
4N-10 & 3N-10 & 2N-10 & N-10 & 10 \\
10N-40 & 10N-40 & 10N-40 & 10N-40 & (N-4) \times \\
\end{bmatrix}

Q-array = \left( \frac{4N-10}{10N-40}, \frac{7N-20}{10N-40}, \frac{9N-30}{10N-40}, (\frac{10N-40}{10N-40})^i \right), i = 1, ..., N-4
$$\text{AR}_D(X) = \begin{cases} 
\frac{1}{N^2} & \frac{1}{10N - 40} \left( 4N - 10 + 7N - 20 + 9N - 30 + 10N - 40 + \frac{10(N - 4)(N - 3)}{2} \right) \\
- \frac{N + 2}{2N} & 
\end{cases}$$

$$= \frac{1}{N^2} \frac{1}{10N - 40} \left( 4N - 10 + 7N - 20 + 9N - 30 + 10N^2 - 70N + 120 + 5N^2 - 35N + 60 \right)$$

$$= \frac{1}{N^2} \frac{1}{10N - 40} \left( 15N^2 - 85N + 120 - \frac{N + 2}{2N} \right)$$

$$= \frac{15N^2 - 85N + 120 - (N + 2)(5N - 20)}{10N(N - 4)} = \frac{15N^2 - 85N + 120 - 5N^2 + 10N + 40}{10N(N - 4)}$$

$$= \frac{10N^2 - 75N + 160}{10N(N - 4)} = \frac{2N^2 - 15N + 32}{2N(N - 4)}$$

**Case II. An example in citation analysis**

One can consider a citing article to be in a dominating position with respect to the article which receives a citation. This idea is now illustrated with an article that has four references and which receives step by step more and more citations, making the structure more and more top-heavy. The different steps are shown in Table 1.

**Table 1. A short citation history of an article with four references**

<table>
<thead>
<tr>
<th>Number of citing articles</th>
<th>N</th>
<th>Array</th>
<th>Value of global D-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
<td>(4,-1,-1,-1,-1)</td>
<td>0.8</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>(9,3,-3,-3,-3)</td>
<td>0.708</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>(9,9,2,-5,-5,-5,-5)</td>
<td>0.621</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>(9,9,9,9,9,0,-9,-9,-9,-9)</td>
<td>0.5</td>
</tr>
<tr>
<td>8</td>
<td>13</td>
<td>(9,9,9,9,9,9,9,9,9,9,9,9,9,9,-4,-17,-17,-17,-17)</td>
<td>0.357</td>
</tr>
</tbody>
</table>
It is possible to derive a general formula for the global D-measure in the case that the original article has \( n \) references and received \( m \) citations, hence \( N = n + m + 1 \). The corresponding array is:

\[
X = \begin{pmatrix}
1, 2, \ldots, 1, 2, \ldots, (1, 2), \ldots, (1, 2) \\
n \text{times} \\
m \times n \times n \times m \times m \times m
\end{pmatrix}
\]

If \( m \geq n \) then the \(|A|\)-array is:

\[
\begin{pmatrix}
1, 2, \ldots, 1, 2, \\
m \times n \times n \times m \times m \times m
\end{pmatrix}
\]

and the Q-array becomes:

\[
AR_d(X) = \frac{1}{N} \left( \sum_{i=1}^{N} q_i \right) - \frac{N + 2}{2N}
\]

One can easily check that, if \( m = n \), the value of \( AR_d(X) \) is equal to 0.5. If \( n = 0 \), corresponding with an article without references, then the result is \( 1/(m+1) \). This clearly is a decreasing power structure for increasing \( m \).

If now \( m \leq n \) then the positive part is equal to \( n(2m+1) \) and we have:

\[
AR_d(X) = \frac{1}{N} \left( \sum_{i=1}^{N} q_i \right) - \frac{N + 2}{2N}
\]

One can easily check that, if \( m = n \), the value of \( AR_d(X) \) is equal to 0.5. If \( n = 0 \), corresponding with an article without references, then the result is \( 1/(m+1) \). This clearly is a decreasing power structure for increasing \( m \).
\[
\frac{1}{m+n+1} \left( \frac{m(m+1)(1+2n)}{2n(1+2m)} + 1 + n + \frac{n+1}{2} \right) - \frac{m+n+3}{2(m+n+1)}
\]

\[
= \frac{1}{m+n+1} \left( \frac{m^2(2n+1) + m(6n^2 + 8n + 1) + 3n(n+1)}{2n(1+2m)} \right) - \frac{m+n+3}{2(m+n+1)}.
\]

Also here the expression becomes equal to 0.5 when \(m = n\). If \(m = 0\), corresponding with an article that has not been cited yet, then the result is \(n/(n+1)\), as we already know.

**Discussion and conclusion**

In this article we provided some further examples of dominance structure in a directed acyclic network. Then we calculated the D-measure when nodes are added to an existing simple network. This was done moving “from a hand to a line” and for a citation network consisting of the ego (the original article) with \(n\) references and \(m\) received citations. When \(n=m\) the network is symmetric and the D-measure is 0.5; when the article did not yet receive a citation the D-measure is \(n/(n+1)\). When the article has no references then the D-measure is \(1/(m+1)\). We consider our contribution as an aspect of the “hardening” of the social sciences.

**Acknowledgements**

The authors thank Leo Egghe (University Hasselt, Belgium) for helpful discussions. This work was supported by NSSFC via 71173154, National Social Sciences Foundation of China.

**References**


An analysis of international research collaboration based on research project data

Lili Yuan ¹ Yanni Hao ² Dengsheng Wu ³* Minglu Li ⁴ Jianping Li ⁵

¹ sonyayll@163.com
Institute of Policy and Management, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing (China)

² haoyn@nsfc.gov.cn
Information Centre, National Natural Science Foundation of China, Beijing (China)

³ wds@casipm.ac.cn (Corresponding author)
Institute of Policy and Management, Chinese Academy of Sciences, Beijing (China)

⁴ liml@nsfc.gov.cn
Bureau of Planning, National Natural Science Foundation of China, Beijing (China)

⁵ ljp@casipm.ac.cn
Institute of Policy and Management, Chinese Academy of Sciences, Beijing (China)

Abstract
International research collaboration is an important form of Science and Technology (S&T) research and is attracting an increasing attention from researchers. While other studies have examined collaboration using publication and patent data, this paper examines funded research collaborative project data by selecting collaborators affiliated with units outside Mainland China. Among a total of 320,000 National Natural Science Fund of China (NSFC) projects conducted from 2006 to 2016, 17,624 projects involving international research collaboration were extracted and included in our empirical analysis. Using the bibliometric method, this paper quantitatively analyses three aspects of the collaboration currently happening in Mainland China: collaboration activity, relative research efforts, and subject field of collaboration. The results demonstrate that among all 78 countries and regions of collaboration, the top fifteen account for
94.73% of all projects on record. The USA is the most important collaboration partner, but the number of research collaborations between the US and Mainland China is around the average of the top fifteen partners. Collaborations with Australia and Netherlands show an increasing trend, while collaborations with Japan and Germany decreased from 2006 to 2016. Of eight fields included in this analysis, the highest number of collaboration projects deal with health sciences, information science or management science.

Introduction

As an essential form of S&T research, international research collaboration is vital to a country’s growth, particularly that of developing countries (Finardi and Buratti 2016). Since 1978, when China adopted a policy of reform and opening-up, the government has encouraged research cooperation with many different countries. According to the administrator of the National Natural Science Fund from the Central Government, the NSFC will continue to increase international research funding in its thirteenth Five-Year Plan for the development of the National Natural Science Fund. International research collaborations enable researchers to share scientific knowledge, create networks of scholarly communication and generate new ideas; they also lower research costs and increase productivity (Katz and Martin 1997; Beaver 2001). Therefore, the study of international research collaboration in China is becoming more and more significant. A better understanding of the existing collaborations will make it possible to make specific recommendations about the establishment of collaborative relationships with other countries in the future.

International research collaboration in China has attracted attention from many researchers. Existing studies mainly focus on specific research collaboration partners or collaboration in specific fields. Some studies examine Chinese research collaborations with specific countries, such as the USA (Wagner et al. 2015), the UK (Zhou et al. 2013) or Germany (Zhou and Bornmann 2015), or with groups of countries, such as BRICS (Finardi 2015; Finardi and Buratti 2016) or G7 countries (He 2009). These partner studies analyse overall collaboration on the macro level. These studies usually conclude that partner countries
have different distributions of collaboration output numbers and collaborate on different subject fields, and the growth trends in collaboration with different countries vary over the given time span. However, some studies on collaboration in some specific areas take a focused micro approach. Generally, the traditional basic disciplines such as mathematics (Zhou and Tian 2014), physics (Zhou and Lv 2015), food and agriculture (Zhou et al. 2013), and cutting-edge technologies like healthcare science and services (Chen et al. 2016) and nanotechnology (Wang et al. 2012; Tang and Shapira 2011) have received the most attention. Most studies examine specific growth, the strength of collaborations, amount of collaboration activity and citation performance. Most studies have concluded that a gap exists between China and other countries and regions in these specific areas.

Most previous studies have used a bibliographic method to study collaboration via publication data (Coccia and Wang 2016), patent data (Guan and He 2007) or a combination of both (Zheng et al. 2012). Publication data is usually extracted from several journal indexes: Thomson Reuters’ Science Citation Index (Expanded)/Social Science Citation Index (SCI/SSCI) (Zhou and Lv 2015; Barth et al. 2014; Leydesdorff et al. 2014), which is the most commonly cited dataset; Elsevier Scopus (Finardi and Buratti 2016); Journal Citation Reports (Yan and Guns 2014) and other minor journal indexes. Patent data is usually retrieved from the United States Patent and Trademark Office (USPTO) database.

Publication data and patent data are the most commonly used data because they are readily available. Many meaningful, extensive conclusions have been reached using these data sources. However, the selection of collaboration databases from different sources is just a method for analysing collaboration from different perspectives. Project collaboration is another kind of collaboration, but it has only been included with other types in collaboration studies. This type of data has many special features which make collaborative project data appropriate for analysing collaboration.

1) Firstly, it is a feasible and reasonable way to analyse collaboration. This is because collaboration project data is itself a kind of direct collaboration, and it contains information not only on the project applicant but the
participants.

2) Secondly, this data reflects collaborations in real time, with no time lag. This is because the collaboration is created the moment the collaborative project is approved. It commonly takes several months to one year or more for articles to be published and patents to be approved.

3) Finally, this type of data enables a more comprehensive analysis of collaboration. This is because collaborative projects result in diverse types of output, so the study is not limited to publications and/or patents.

This paper combines data about collaborative projects in Mainland China with that of other countries and regions to analyse the current status of international research collaboration in Mainland China. This empirical analyses uses the NSFC data on funded international collaborative projects. Using the bibliometric method, we quantitatively analyse three aspects of current collaboration: collaborative activity, relative strength of collaboration and collaboration research efforts in each field. Based on this analysis and a discussion of it, some conclusions are reached on the international collaboration between Mainland China and other countries and regions.

**Data and Methods**

*Data sources*

The data used in this paper was extracted from the NSFC dataset of funded projects. From a total of 0.32 million NSFC projects from 2006 to 2016, 17,624 projects involving international research collaboration were extracted as our research dataset. The extracted dataset includes the personnel and project information. The dataset includes 17,624 collaborative projects with 22,655 persons from 78 countries and regions (including Hong Kong, Macau and Taiwan). Each project involves one applicant and one or more participants. This is why there are more people than projects. Each project includes people from different countries or regions (including Hong Kong, Macau and Taiwan) in addition to Mainland China. Table 1 shows the sample for one project.
People from Hong Kong, Macau and Taiwan are included in the datasets. As Special Administration Regions of China, Hong Kong and Macau are much more international than Mainland China (Wang et al. 2013). For historical reasons, Taiwan also differs from Mainland China in many aspects, such as education and culture.

Data processing
Before analysis, the data was processed using the following steps.

Firstly, the data was cleaned. This includes filtering out records with null fields, uniquely identifying each person and implementing name disambiguation for each person’s unit or units of affiliation. For detailed information on name disambiguation, please refer to Yan and Sugimoto (2011).

Secondly, the affiliated units were allocated to the countries or regions where they are belonged to.

Finally, the collaborative projects were assigned to one of eight subject fields according to the area of application. The subject field was divided into eight categories corresponding to the eight scientific departments at NSFC. The eight subject fields are Mathematical and Physical Sciences (MPS), Chemical Sciences (CS), Life Sciences (LS), Earth Sciences (ES), Engineering and Materials Sciences (EMS), Information Sciences (IS), Management Sciences (MS) and Health Sciences (HS). Moreover, some special projects, such as Emergency projects and Director Fund projects, did not fit into any of the eight subject fields due to their specificity. These were categorized as Other and will not be analysed in this paper. Of a total of 17,624 collaboration projects, 17,503 projects were allocated to one of the eight subject fields, and the remaining 121 projects were classified as Other.

<table>
<thead>
<tr>
<th>No.</th>
<th>Person’s name</th>
<th>Affiliation</th>
<th>Project name</th>
<th>Application code</th>
<th>Year funded</th>
<th>Role in the project</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Kevin J.K.</td>
<td>Oxford University</td>
<td>Modelling of XXX</td>
<td>A0113</td>
<td>2016</td>
<td>participant</td>
</tr>
<tr>
<td>2</td>
<td>Paul Cooper</td>
<td>UCLA</td>
<td>Investigation of brain structure and XXX</td>
<td>H3101</td>
<td>2015</td>
<td>applicant</td>
</tr>
</tbody>
</table>
Results and Discussion

This section addresses three aspects of international research collaboration between Mainland China and other countries all over the world: collaborative activity, relative research effort and subject field of collaboration.

Collaborative activity

This section provides an overview of the collaboration between Mainland China and other countries and regions. It shows the total absolute collaboration numbers and the growth in collaborative projects from 2006 to 2016.

Table 2 shows detailed information about the total number of persons and projects in specific countries and regions. The top ten countries and regions account for 90.48% of the total 17,624 collaboration projects, and the top fifteen account for 94.73%. The number of people and projects collaborating with researchers from Mainland China varies widely from country to country. Forty-five point sixty-seven percent of the collaboration projects in Mainland China involve collaboration with the USA. Even within the top fifteen, there is a huge gap between countries in the share of international collaboration projects. South Korea, Macau and Denmark are each involved in less than 1% of the total projects. We conclude that nearly 95% of international research projects involve collaboration with the top fifteen countries and regions, and almost half of the collaborations are with the USA.

Table 2. Numbers of projects and persons in top fifteen countries and regions collaborating with Mainland China

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country or region</th>
<th>Funded projects</th>
<th>Related persons</th>
<th>% Funded projects</th>
<th>Cumulative projects</th>
<th>% Cumulative projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>USA</td>
<td>8505</td>
<td>9473</td>
<td>45.67%</td>
<td>8505</td>
<td>45.67%</td>
</tr>
<tr>
<td>2</td>
<td>Hong Kong</td>
<td>2180</td>
<td>4154</td>
<td>11.71%</td>
<td>10685</td>
<td>57.37%</td>
</tr>
<tr>
<td>3</td>
<td>UK</td>
<td>1322</td>
<td>1486</td>
<td>7.10%</td>
<td>12007</td>
<td>64.47%</td>
</tr>
<tr>
<td>4</td>
<td>Australia</td>
<td>1054</td>
<td>1179</td>
<td>5.66%</td>
<td>13061</td>
<td>70.13%</td>
</tr>
<tr>
<td>5</td>
<td>Japan</td>
<td>1009</td>
<td>1070</td>
<td>5.42%</td>
<td>14070</td>
<td>75.55%</td>
</tr>
<tr>
<td>6</td>
<td>Singapore</td>
<td>660</td>
<td>752</td>
<td>3.54%</td>
<td>15524</td>
<td>83.35%</td>
</tr>
<tr>
<td>7</td>
<td>Germany</td>
<td>569</td>
<td>653</td>
<td>3.06%</td>
<td>16093</td>
<td>86.41%</td>
</tr>
<tr>
<td>8</td>
<td>France</td>
<td>405</td>
<td>457</td>
<td>2.17%</td>
<td>16498</td>
<td>88.58%</td>
</tr>
<tr>
<td>9</td>
<td>Taiwan</td>
<td>353</td>
<td>448</td>
<td>1.90%</td>
<td>16851</td>
<td>90.48%</td>
</tr>
<tr>
<td>10</td>
<td>Netherlands</td>
<td>208</td>
<td>225</td>
<td>1.12%</td>
<td>17059</td>
<td>91.60%</td>
</tr>
</tbody>
</table>
To clearly illustrate the profile of international research collaboration between Mainland China and the top fifteen countries and regions, Table 3 shows the number of collaborative projects in each country for each year from 2006 to 2016. Before the results are discussed, the application limit issued in 2012 by the NSFC on General Projects should be mentioned. This limit forbade re-applications by researchers who had applied for but not received General Project funding in the past two consecutive years. This application limit affected the number of the project applications, as reflected in the numbers shown in Table 3. In every year, more projects involved collaboration with the USA than any other country, and collaboration with the USA grew rapidly before 2012. The number of collaborations with the USA reached its maximum of 1,123 in 2012 and has stayed over 1,000 since then, despite a slight decrease. Collaboration with Hong Kong experienced a similar upward trend from 2009 to 2012, but decreased rapidly in 2014. The UK and Australia witnessed a steady increase from 2006 to 2016. A similar situation happened in Canada and Japan; collaborations with both countries grew steadily before 2012 and then decreased slightly. Overall, 2012 is a turning point for trends in collaborative projects. From 2006 to 2012, collaborations with nearly all of the top fifteen grew, while collaborations with over half of the top fifteen decreased after 2012. Although the absolute number of other countries and regions collaborating with Mainland China is far smaller than that of the USA, the number of collaborations still show an upward trend, which means that China has a propensity to collaborate.

Table 3. Growth in the number of collaborative projects with Mainland China from 2006 to 2016

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>299</td>
<td>338</td>
<td>446</td>
<td>465</td>
<td>675</td>
<td>926</td>
<td>1123</td>
<td>1078</td>
<td>1032</td>
<td>1061</td>
<td>1062</td>
<td></td>
</tr>
<tr>
<td>Hong Kong</td>
<td>87</td>
<td>86</td>
<td>143</td>
<td>127</td>
<td>181</td>
<td>249</td>
<td>314</td>
<td>296</td>
<td>212</td>
<td>241</td>
<td>244</td>
<td></td>
</tr>
</tbody>
</table>
Relative research effort

This section analyses the relative research effort of specific countries and regions collaborating with Mainland China. The absolute number of collaboration projects between Mainland China and a specific country or region depends on the country’s size and its propensity to collaborate (Luukkonen et al. 1993). The Activity Index (AI) is a relative indicator which was first proposed by Frame (1977) and was used by Bujdoso and Braun (1983). The AI was later defined in detail by Schubert and Braun (1986) as characterizing the relative effort a country devotes to a given subject field (Schubert and Braun 1986; Kumar and Garg 2005). It can eliminate the effect of country size when comparing the absolute number of collaborative projects. Many researchers have modified and developed this definition (Garg and Padhi 2001; Kumar and Garg 2005; Garg et al. 2006; He 2009).
Here, \( AI = (P_{ij} / P_{ii}) / (P_{cj} / P_{ct}) \)

where \( P_{ij} \) is the number of collaborative projects between Mainland China and country or region \( i \) in the given year \( j \); \( P_{ii} \) is the number of projects between Mainland China and country or region \( i \) from 2006 to 2016; \( P_{cj} \) is the number of projects between Mainland China and all countries and regions \( c \) in the given year \( j \); \( P_{ct} \) is the number of projects between Mainland China and all countries and regions \( c \) from 2006 to 2016; and \( i = 1, 2, \ldots, 15; j = 2006, 2007, \ldots, 2016 \).

If \( AI = 1 \), the country or region’s research effort in the given year corresponds to the average effort of the top fifteen countries and regions; \( AI > 1 \) reflects an above-average effort; \( AI < 1 \) indicates a below-average effort.

In this way, the relative effort of each country or region in any year can be compared. For example, we can compare the USA’s relative research effort in 2007 to that of Japan in 2015.

Figure 1 shows the AI distribution profiles for the top fifteen countries and regions collaborating with Mainland China. To
clearly show the distribution profiles, the top fifteen are divided into four parts, (a) (b) (c) (d), in Figure 1, according to their similar or adverse variation characteristics over time. Figure 1(a) shows that research efforts toward collaboration between Mainland China and the USA, Hong Kong, the UK or Canada were near the average of the top fifteen countries from 2006 to 2016 (AI = 1). After diminishing the effect of country size, research efforts for these countries decrease to the average. Although these four countries and regions among the top five in absolute number of collaborative projects with Mainland China, they seem to show no preference for research collaboration with Mainland China.

Figure 1(b) shows two different growth trends. The research efforts of Mainland China with Japan and Germany decreased steadily, from around 1.5 to 0.7, and collaboration efforts with Australia and The Netherlands increased gradually, from 0.5 to nearly 1.5. This means that Australia and The Netherlands demonstrate a growing desire for collaboration, while Japan and Germany demonstrate an opposite trend. The reason for this should be determined, and policies should be established to provide these countries with more incentive for collaboration. These policies could include financial grants, granting more collaborative project, and providing more support for exchange students.

In contrast to Figures 1(a) and 1(b), Figures 1(c) and 1(d) display an irregular, jumpy variations over time. The difference between Figure 1(c) and Figure 1(d) is the fluctuations in the size of the variations. In Figure 1(c), AI varies less than in 1(d).

To summarize, from 2006 to 2016, research effort in the USA, Hong Kong, the UK and Canada is around the average of that of the top fifteen. Australia and The Netherlands show an overall rising trend in research collaboration efforts with Mainland China, and collaborations with Japan and Germany are decreasing. The other countries and regions in the top fifteen show random variations in collaborative efforts.

**Collaboration field**

In the first two sections, all projects are compared without considering the subject fields they address. In fact, after many
years of development, there are significant differences and imbalances between subject fields in different countries and regions. However, an analysis of collaboration subject fields in different countries and regions is critical for a thorough understanding of collaborative projects in Mainland China. Therefore, Table 4 shows the number of collaborative projects in each subject field in the top fifteen countries and regions.

As the bottom row of Table 4 demonstrates, the absolute number of collaborations in each subject field is quite different. Of 17,624 projects examined, the most projects (4,293) are in the field of HS, and the fewest (566) are in CS. The most common fields for collaborative research are HS, IS and MS. MS accounts for 15.98% of total projects in this study and is the third most common field. However, of the total 0.32 million NSFC projects from 2006 to 2016, only 4.38% deal with MS, and it is the least common field of research overall. This big difference in total and collaborative MS studies means that this field is quite popular for international research and shows a kind of propensity.

The USA has the most collaborative projects in all subject fields. The bar chart in the last column in Table 4 illustrates the wide variation in the number of collaborative projects between Mainland China and the top fifteen in different subject fields. There is a great deal of variation and imbalance in the numbers of collaborations in different subject fields in all countries and regions. HS, IS and MS are the most common fields for collaboration, sharing nearly 60% of all collaborative projects. Moreover, there is a great deal of international research in MS. This kind of difference and unbalance may be partly caused by the dominance of certain disciplines in some countries, such as HS in the USA.
Table 4. Numbers of collaborative projects in each field between the top fifteen and Mainland China

<table>
<thead>
<tr>
<th>Country or Region</th>
<th>MPS</th>
<th>CS</th>
<th>LS</th>
<th>ES</th>
<th>EMS</th>
<th>IS</th>
<th>MS</th>
<th>HIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>582</td>
<td>233</td>
<td>1036</td>
<td>735</td>
<td>662</td>
<td>1149</td>
<td>1202</td>
<td>2880</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>210</td>
<td>110</td>
<td>143</td>
<td>162</td>
<td>215</td>
<td>455</td>
<td>532</td>
<td>335</td>
</tr>
<tr>
<td>UK</td>
<td>85</td>
<td>36</td>
<td>119</td>
<td>151</td>
<td>225</td>
<td>284</td>
<td>240</td>
<td>177</td>
</tr>
<tr>
<td>Australia</td>
<td>76</td>
<td>28</td>
<td>83</td>
<td>132</td>
<td>169</td>
<td>237</td>
<td>187</td>
<td>139</td>
</tr>
<tr>
<td>Canada</td>
<td>73</td>
<td>31</td>
<td>116</td>
<td>128</td>
<td>80</td>
<td>188</td>
<td>180</td>
<td>206</td>
</tr>
<tr>
<td>Japan</td>
<td>53</td>
<td>32</td>
<td>103</td>
<td>95</td>
<td>172</td>
<td>121</td>
<td>62</td>
<td>154</td>
</tr>
<tr>
<td>Singapore</td>
<td>66</td>
<td>30</td>
<td>33</td>
<td>12</td>
<td>86</td>
<td>233</td>
<td>133</td>
<td>65</td>
</tr>
<tr>
<td>Germany</td>
<td>57</td>
<td>26</td>
<td>89</td>
<td>131</td>
<td>58</td>
<td>59</td>
<td>47</td>
<td>100</td>
</tr>
<tr>
<td>France</td>
<td>57</td>
<td>4</td>
<td>37</td>
<td>60</td>
<td>41</td>
<td>124</td>
<td>46</td>
<td>34</td>
</tr>
<tr>
<td>Taiwan</td>
<td>24</td>
<td>6</td>
<td>35</td>
<td>65</td>
<td>26</td>
<td>39</td>
<td>66</td>
<td>25</td>
</tr>
<tr>
<td>Netherlands</td>
<td>14</td>
<td>4</td>
<td>33</td>
<td>25</td>
<td>22</td>
<td>28</td>
<td>39</td>
<td>42</td>
</tr>
<tr>
<td>Sweden</td>
<td>10</td>
<td>10</td>
<td>22</td>
<td>16</td>
<td>20</td>
<td>26</td>
<td>22</td>
<td>69</td>
</tr>
<tr>
<td>Macau</td>
<td>16</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>10</td>
<td>31</td>
<td>22</td>
<td>32</td>
</tr>
<tr>
<td>South Korea</td>
<td>18</td>
<td>8</td>
<td>19</td>
<td>9</td>
<td>21</td>
<td>40</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>Denmark</td>
<td>5</td>
<td>6</td>
<td>19</td>
<td>10</td>
<td>18</td>
<td>11</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

Total in 17,503 projects 1346 566 1893 1735 1825 3025 2820 4295
%Total in 17,503 projects 7.69% 3.23% 10.82% 9.91% 10.43% 17.28% 16.11% 24.55%

Total in 0.32 million projects 36326 31268 51180 30729 52225 35455 14124 69500
%Total in 0.32 million projects 11.32% 9.75% 15.95% 9.58% 16.28% 11.05% 4.40% 21.66%
A comparison of the absolute number of collaborative projects does not allow for differences in country size or the propensity of a country or region when collaborating with Mainland China. Hence, the International Collaboration Index (ICI) is adopted to measure the relative research effort of the top fifteen in each field. ICI is an improved internationalization index and was defined by Garg and Padhi (2001).

Here, \( ICI = \left( \frac{P_{ij}}{P_{it}} \right) / \left( \frac{P_{cj}}{P_{ct}} \right) \)

where \( P_{ij} \) is the number of projects on which country or region \( i \) collaborates with Mainland China in the given subject field \( j \); \( P_{it} \) is the number of projects on which country or region \( i \) collaborates with Mainland China in all subject fields; \( P_{cj} \) is the number of projects on which all the countries and regions \( c \) collaborate with Mainland China in the given subject field \( j \); and \( P_{ct} \) is the number of projects on which all the countries and regions \( c \) collaborate with Mainland China in all subject fields.

For the eight subject fields, Table 5 shows the highest and lowest ICI and the corresponding countries and regions collaborating with Mainland China for eight subject fields. Denmark ranks first in CS and LS but last in MPS and IS. France ranks first in MPS and last in CS and HS. In the field of EMS, the USA has the lowest effort with Mainland China, while Japan has the highest effort. Hong Kong has the highest collaboration effort in MS, and Japan has the lowest effort. The biggest effort gap occurred in Singapore; of eight fields, the lowest effort is 0.18, for ES, and the highest is 2.05, for IS. It can be concluded that no one country or region ranks highest or lowest in all subject fields. The top fifteen countries and regions vary in each research field, and there are no obvious trends. Contrasted with the absolute advantage in the number of collaborative projects with the USA, the USA’s research efforts are below the average in six fields, and its effort in EMS is the lowest effort with Mainland China.
Table 5. Highest and lowest ICI for collaborations with Mainland China and the corresponding countries and regions in eight subject fields

<table>
<thead>
<tr>
<th>Subject field</th>
<th>Highest effort</th>
<th>Lowest effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Country or region</td>
<td>Research effort</td>
</tr>
<tr>
<td>MPS</td>
<td>France</td>
<td>1.84</td>
</tr>
<tr>
<td>CS</td>
<td>Denmark</td>
<td>1.70</td>
</tr>
<tr>
<td>LS</td>
<td>Denmark</td>
<td>1.61</td>
</tr>
<tr>
<td>ES</td>
<td>Germany</td>
<td>2.33</td>
</tr>
<tr>
<td>EMS</td>
<td>Japan</td>
<td>2.08</td>
</tr>
<tr>
<td>IS</td>
<td>Singapore</td>
<td>2.05</td>
</tr>
<tr>
<td>MS</td>
<td>Hong Kong</td>
<td>1.53</td>
</tr>
<tr>
<td>HS</td>
<td>Sweden</td>
<td>1.44</td>
</tr>
</tbody>
</table>

Conclusions
Mainland China has established international research collaboration with many countries and regions in the past 40 years, and it has continuously attached great importance to international research collaboration. Therefore, it increasingly important to have an overall understanding of existing collaboration. This paper analyses three aspects of international collaboration with Mainland China: collaboration activity, relative research effort and subject field of collaboration. Several conclusions can be drawn.

1) Of the 78 collaborative countries and regions in this study, the top fifteen account for 94.73% of the total 17,624 collaborative projects. The USA has the most collaborative projects with Mainland China. A trend turning point occurred in 2012. Collaborations with nearly all of the top fifteen increased before 2012, and collaborations decreased with over half of the top fifteen after 2012. Although the absolute number of other countries and regions collaborating with Mainland China is far less than that of the USA, they still show an upward trend, which indicates a propensity to collaborate.

2) From 2006 to 2016, the research effort of the USA, Hong Kong, the UK and Canada is around the average of the top
fifteen (AI=1). Collaboration effort between Mainland China and Australia and The Netherlands shows an overall upward trend; Japan and Germany show a downward trend. Other countries and regions show random variation in collaboration effort.

3) There are significant differences across the eight subject fields, and there are no obvious trends in different countries and regions in subject fields. HS, IS and MS are the top three fields, accounting for nearly 60% of all collaborative projects. MS is a popular subject for international research collaboration. No one country or region ranks the highest or lowest in all subject fields. The USA ranks first in total numbers in all these fields, but the USA’s collaboration efforts are below average in six fields, and the USA’s collaborative effort in EMS is the lowest effort with Mainland China.

Future studies on collaboration should include different data sources, such as publication data and patent data (which are widely used) as well as applied project data. The integration of different data sources would provide a more comprehensive understanding of international research collaboration with Mainland China.

Acknowledgments
This research has been supported by grants from the National Natural Science Foundation of China (71425002, 71640007), the Special Fund Project of the Qinghai Province for the Transformation of Scientific and Technological Achievements (2016-GX-109), and the Youth Innovation Promotion Association of the Chinese Academy of Sciences (2013112).

References


The Evolution of China’s Role in the International Scientific Collaboration Network

Zhihui Zhang\textsuperscript{1}, Jason E. Rollins\textsuperscript{2}, Evangelia Lipitakis\textsuperscript{3}

\textsuperscript{1}zhihui.zhang@clarivate.com  
Clarivate Analytics (China)

\textsuperscript{2}jason.rollins@clarivate.com  
Clarivate Analytics (United States)

\textsuperscript{3}evangelia.lipitakis@clarivate.com  
Clarivate Analytics (United Kingdom)

Abstract
In this paper we analyze the evolution of China’s growing importance in international scientific collaboration over the past decade. Using data from nearly 12 million articles indexed in Clarivate Analytics's Web of Science Core Collection we develop novel weighted and unweighted centrality measures to quantify China’s emerging role in the global network of scientific collaboration. We analyze the networks formed by international co-authorship in two five-year periods: 2006-2010, and 2011-2015. This analysis highlights China’s sharp increase in prominence in international scientific collaborations. Our findings help to enumerate China’s position as the fastest growing international collaborator among the 40 countries studied. We further contextualize our work in a discussion of international scientific collaboration as both a key driver of China’s economy and its emerging perception as a first-world innovator and intellectual power. Finally, we suggest directions for further research including more granular analysis by county and academic discipline.

Conference Topic
International scientific collaboration; Country-level studies;

Introduction
China’s impressive economic development over the past several decades (resulting from the "reform and opening-up" polices begun in the late 1970's) has been widely lauded by a range of observers (Engardio, 2005; Zhou, 2015). The growing prominence of China in global science and innovation is also becoming a subject of broad fascination and is studied ever more frequently (Engardio, 2005; Klochikhin, 2013; Zhou, 2015). Research often drives innovation and China has heavily invested in research in recent years, particularly in the past decade, with substantial results. We see a symbiotic connection between China’s emerging scientific and economic significance and one that can be validated by quantifying the growth in international scientific collaborations, as measured through co-authored scientific papers, between China and other leading countries with mature, knowledge-based economies. Based on papers indexed in Clarivate Analytics’s Web of Science Core Collection, China’s research output surpassed the UK in 2008 and is now second only to the United States. Certainly from the perspective of publication volume, China is playing a more and more important role in global scientific research. This inevitably also has positive economic impact, as many of the breakthroughs
presented in scientific publications will evolve into commercial technologies and products driving investment and employment growth for China.

Collaborative research can inspire participants by exchanging and synthesizing diverse ideas and viewpoints resulting in high impact science. Researchers from different countries are varied in culture and educational background, and often have distinct views on the same research problem. Furthermore, international collaboration can integrate and leverage both physical and intellectual resources from participating countries; this is a clear advantage for research in the era of big science (Niu & Qiu, 2014). This is also important in explaining why the citation impact of publications with international co-authorship is statistically higher than domestic-only publications (Frenken, Hardeman, & Hoekman, 2009; Glänzel & Schubert, 2001; Narin, Stevens, & Whitlow, 1991; Nomaler, Frenken, & Heimeriks, 2013). These generally mutually beneficial results explain why most countries show interest in collaborating with other countries in scientific research.

When considering the role of countries in the scientific collaboration network, the distinctions are remarkable. Some leading nations like the US, the UK, Germany, and France have researchers who co-authored a significant number of papers with many other countries scattered around the world; conversely, there are also examples of countries that only have intensive collaborations with their local geographic neighbors. From this perspective the US, the UK, Germany, and France are the key nodes of the global collaboration network. Over the past several years the collaboration among countries in scientific research has intensified considerably (Zhou & Glänzel, 2010). With China’s rapidly increasing volume of publications, it is emerging as an important collaborator with other countries in scientific research. Academics studying scientometrics and research policy, the Chinese government, research administrators, and even the general public are all interested in China’s emerging position in the international scientific collaboration network. This paper endeavors to quantify the prominence of countries in the collaboration network and investigate the evolution of China’s role over the past decade.

To measure the role and importance of nodes in a network, the most commonly used methods are from Social Network Analysis (SNA). Degree centrality, closeness centrality, and betweenness centrality are three commonly used measures in SNA (Freeman, 1978; Hanneman & Riddle, 2005). However, these measures cannot be reliably applied to the analysis of the international collaboration network of countries since such a network is usually a complete graph. A network is seen as a complete graph if any two nodes have a direct connection, or there is an edge between any two nodes (Wilson, 2012). In a collaboration network formed by countries whose publication output is beyond very minimal, any two countries would have a collaboration relationship, and such a network is a complete graph. In a complete graph, the values of centrality measures for all the nodes will not be significantly different. It is therefore inappropriate to gauge the roles of different nodes using traditional centrality measures as they would not be distinguishable.
International scientific collaboration is a complex and heterogeneous phenomenon, and bibliometric methods can offer deep insight into national characteristics in international co-authorship relations (Glänzel, 2001). In particular, International co-authorship has emerged as a standard proxy of international collaboration (Hennemann, Wang, & Liefner, 2011; Katz & Martin, 1997).

Methods and Data
We use two indicators to measure the role and importance of countries in the global collaboration network. We look at countries as aggregate proxies for all of the papers co-authored with at least one international author. The core idea underlying the indicators is based on the fact that for two countries \( i \) and \( j \), if the collaborated publications between \( i \) and \( j \) take up a high proportion of all collaborated publications of \( j \), then \( i \) is an important collaborator of \( j \). Let \( C_{ij} \) denotes the number of collaborated publications between countries \( i \) and \( j \), and \( C_j \) denotes the total number of publications with international co-authorship of \( j \). The ratio \( C_{ij}/C_j \) measures the importance of \( i \) as a collaborator of \( j \) among all \( j \)'s collaborators (Zhou & Glänzel, 2010; Zitt, Bassecoulard, & Okubo, 2000). The ratio is the numerator of co-authorship preference index proposed by Glänzel and Schubert (Archambault, Beauchesne, Côté, & Roberge, 2011; Schubert & Glänzel, 2006). Formula (1) accumulates \( i \)'s importance among all its collaborators and represents \( i \)'s overall importance as a collaborator within the network.

\[
\sum_{j \neq i} \frac{C_{ij}}{C_j}
\]

Formula (1)

In formula (1) we do not take the volume of \( j \)'s publications into account when we accumulate \( i \)'s importance over \( j \). For a large country \( j \) with a high volume of publications and a small country \( k \) with small number of publications, \( C_{ij}/C_j \) may be much lower than \( C_{ik}/C_k \). But \( j \) is usually more influential than \( k \) since \( j \) typically publishes many more papers and attracts many more citations. In this sense formula (1) does not consider the impact of country \( j \) when we evaluate \( i \)'s importance. To overcome the limitations of formula (1) we can weight the ratio \( C_{ij}/C_j \) by the global share of \( j \)'s publications \( P_j \). \( P_j \) is calculated as the number of \( j \)'s publications divided by the number of world’s total publications.

\[
\sum_{j \neq i} \frac{C_{ij}}{C_j} P_j
\]

Formula (2)

In this paper both indicators will be used to measure the roles of countries in the collaboration network.

The data used in our analysis comprises papers indexed in Clarivate Analytics’s Web of Science Core Collection and published between 2006 and 2015. Only papers of document type Article and Review are considered.
To construct and analyze the international scientific collaboration network in a realistic way, only the 40 largest countries (measured by overall publication output between 2006 and 2015) are considered. Researchers from the 40 countries published 11,947,664 papers in all. Note that if a paper is co-authored by scientists from multiple countries, this paper is only counted once. The total volume of papers published by the 40 countries accounts for 95.4% of the world’s output. So, at least for the purposes of our present research, the collaboration network formed by these 40 countries can reasonably represent the entire international collaboration network well. Plus, it is more convenient for us to calculate the countries’ performance from formula (1) and (2) in this tailored network, especially given the space restraints for this paper. The measures from formula (1) and (2) are called unweighted and weighted centrality respectively.

**Results**

Table 1 shows the unweighted centralities of the 40 countries with the highest publication output between 2006 and 2015. The rank is based on the total number of publications (both internationally co-authored and domestic-only collaborations are taken into account.) Perhaps not unexpectedly, the centrality of the US is much higher than that of all other countries studied. This suggests that most of the world’s countries have substantial collaborations with the US and demonstrates the core role of the US in the global collaboration network. The UK, Germany, and France also attract substantial collaborations and are important nodes of the network. Clearly, a commonality of the US, the UK, Germany, and France is that they are all well-established countries with a large volume of scientific publication output. China and Japan also produce many publications but their centralities are not as high as their positions in publication output. Although China’s overall publication volume is second only to the US, China’s centrality is just slightly higher than that of the Netherlands, a small country, but one widely known for innovation in science and technology.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Centrality</th>
<th>Rank</th>
<th>Country</th>
<th>Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>USA</td>
<td>12.01</td>
<td>21</td>
<td>Belgium</td>
<td>1.74</td>
</tr>
<tr>
<td>2</td>
<td>China</td>
<td>2.83</td>
<td>22</td>
<td>Denmark</td>
<td>1.41</td>
</tr>
<tr>
<td>3</td>
<td>UK</td>
<td>6.70</td>
<td>23</td>
<td>Israel</td>
<td>0.76</td>
</tr>
<tr>
<td>4</td>
<td>Germany</td>
<td>6.45</td>
<td>24</td>
<td>Austria</td>
<td>1.43</td>
</tr>
<tr>
<td>5</td>
<td>Japan</td>
<td>2.07</td>
<td>25</td>
<td>Finland</td>
<td>1.15</td>
</tr>
<tr>
<td>6</td>
<td>France</td>
<td>4.67</td>
<td>26</td>
<td>Grace</td>
<td>1.06</td>
</tr>
<tr>
<td>7</td>
<td>Canada</td>
<td>2.76</td>
<td>27</td>
<td>Mexico</td>
<td>0.70</td>
</tr>
<tr>
<td>8</td>
<td>Italy</td>
<td>3.76</td>
<td>28</td>
<td>Portugal</td>
<td>1.06</td>
</tr>
<tr>
<td>9</td>
<td>Spain</td>
<td>3.47</td>
<td>29</td>
<td>Norway</td>
<td>1.09</td>
</tr>
<tr>
<td>10</td>
<td>India</td>
<td>1.40</td>
<td>30</td>
<td>Singapore</td>
<td>0.45</td>
</tr>
</tbody>
</table>
For the 20 countries with the highest centralities in Table 1, Table 2 shows the increase of centralities between the two five-year periods 2006-2010 and 2011-2015. Table 2 clearly displays that China has the largest increase of centrality, over the periods studied, with its centrality ranking rising from 10th position in 2006-2010 to 7th place in 2011-2015. This unmistakably indicates that along with a rapid increase in publication volume, China is also emerging as an important collaborator with researchers in other countries around the world.

Table 2. Centrality increases of 20 countries

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>USA</td>
<td>11.87</td>
<td>12.11</td>
<td>0.24</td>
</tr>
<tr>
<td>2</td>
<td>UK</td>
<td>6.26</td>
<td>7.02</td>
<td>0.76</td>
</tr>
<tr>
<td>3</td>
<td>Germany</td>
<td>6.07</td>
<td>6.71</td>
<td>0.64</td>
</tr>
<tr>
<td>4</td>
<td>France</td>
<td>4.34</td>
<td>4.89</td>
<td>0.55</td>
</tr>
<tr>
<td>5</td>
<td>Italy</td>
<td>3.13</td>
<td>4.15</td>
<td>1.02</td>
</tr>
<tr>
<td>6</td>
<td>Spain</td>
<td>2.74</td>
<td>3.94</td>
<td>1.20</td>
</tr>
<tr>
<td>7</td>
<td>China</td>
<td>1.99</td>
<td>3.36</td>
<td>1.37</td>
</tr>
<tr>
<td>8</td>
<td>Canada</td>
<td>2.57</td>
<td>2.88</td>
<td>0.31</td>
</tr>
<tr>
<td>9</td>
<td>Netherlands</td>
<td>2.33</td>
<td>3.04</td>
<td>0.71</td>
</tr>
<tr>
<td>10</td>
<td>Australia</td>
<td>1.99</td>
<td>2.89</td>
<td>0.90</td>
</tr>
<tr>
<td>11</td>
<td>Switzerland</td>
<td>1.86</td>
<td>2.65</td>
<td>0.79</td>
</tr>
<tr>
<td>12</td>
<td>Sweden</td>
<td>1.90</td>
<td>2.37</td>
<td>0.47</td>
</tr>
</tbody>
</table>
Table 3 lists China’s top 10 collaborators between 2011 and 2015. The fourth column $C_{ij}/C_j$ calculates the proportion of China’s international co-authored publications with each country in each country’s total collaborations. Focusing on the values of $C_{ij}/C_j$, China is the least prominent collaborator for France and the Netherlands, and the most important collaborator for Singapore. Papers co-authored by researchers from China and Singapore account for more than 30% of Singapore’s total international collaborations.

Table 3. China’s top 10 collaborators between 2011 and 2015

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Collaborated Papers</th>
<th>$C_{ij}/C_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>USA</td>
<td>122936</td>
<td>18.0%</td>
</tr>
<tr>
<td>2</td>
<td>UK</td>
<td>25567</td>
<td>8.6%</td>
</tr>
<tr>
<td>3</td>
<td>Australia</td>
<td>22667</td>
<td>16.4%</td>
</tr>
<tr>
<td>4</td>
<td>Japan</td>
<td>21672</td>
<td>18.9%</td>
</tr>
<tr>
<td>5</td>
<td>Canada</td>
<td>19234</td>
<td>11.9%</td>
</tr>
<tr>
<td>6</td>
<td>Germany</td>
<td>19230</td>
<td>7.1%</td>
</tr>
<tr>
<td>7</td>
<td>France</td>
<td>11992</td>
<td>6.0%</td>
</tr>
<tr>
<td>8</td>
<td>Singapore</td>
<td>11127</td>
<td>31.1%</td>
</tr>
<tr>
<td>9</td>
<td>South Korea</td>
<td>10714</td>
<td>14.2%</td>
</tr>
<tr>
<td>10</td>
<td>Netherlands</td>
<td>6604</td>
<td>6.0%</td>
</tr>
</tbody>
</table>

Next, we look at formula (2) to calculate the weighted centralities of countries. For readability all the centralities in the following tables are multiplied by 100.

Table 4. Weighted centralities of 40 largest countries

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Centrality</th>
<th>Rank</th>
<th>Country</th>
<th>Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>USA</td>
<td>31.61</td>
<td>21</td>
<td>Belgium</td>
<td>4.34</td>
</tr>
<tr>
<td>Rank</td>
<td>Country</td>
<td>Centrality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-----------</td>
<td>------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>China</td>
<td>10.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>UK</td>
<td>16.97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Germany</td>
<td>16.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Japan</td>
<td>6.87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>France</td>
<td>11.83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Canada</td>
<td>9.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Italy</td>
<td>9.41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Spain</td>
<td>7.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>India</td>
<td>3.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Australia</td>
<td>7.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>South Korea</td>
<td>4.35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Brazil</td>
<td>3.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Netherlands</td>
<td>7.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Russia</td>
<td>3.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Switzerland</td>
<td>6.37</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Turkey</td>
<td>1.84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Sweden</td>
<td>5.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Poland</td>
<td>3.29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Iran</td>
<td>1.23</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>Denmark</td>
<td>3.35</td>
</tr>
<tr>
<td>23</td>
<td>Israel</td>
<td>2.29</td>
</tr>
<tr>
<td>24</td>
<td>Austria</td>
<td>3.36</td>
</tr>
<tr>
<td>25</td>
<td>Finland</td>
<td>2.62</td>
</tr>
<tr>
<td>26</td>
<td>Grace</td>
<td>2.36</td>
</tr>
<tr>
<td>27</td>
<td>Mexico</td>
<td>1.83</td>
</tr>
<tr>
<td>28</td>
<td>Portugal</td>
<td>2.37</td>
</tr>
<tr>
<td>29</td>
<td>Norway</td>
<td>2.51</td>
</tr>
<tr>
<td>30</td>
<td>Singapore</td>
<td>1.82</td>
</tr>
<tr>
<td>31</td>
<td>Czech</td>
<td>2.34</td>
</tr>
<tr>
<td>32</td>
<td>South Africa</td>
<td>1.77</td>
</tr>
<tr>
<td>33</td>
<td>New Zealand</td>
<td>1.60</td>
</tr>
<tr>
<td>34</td>
<td>Argentina</td>
<td>1.42</td>
</tr>
<tr>
<td>35</td>
<td>Ireland</td>
<td>1.62</td>
</tr>
<tr>
<td>36</td>
<td>Malaysia</td>
<td>1.03</td>
</tr>
<tr>
<td>37</td>
<td>Romania</td>
<td>1.29</td>
</tr>
<tr>
<td>38</td>
<td>Egypt</td>
<td>0.97</td>
</tr>
<tr>
<td>39</td>
<td>Hungary</td>
<td>1.77</td>
</tr>
<tr>
<td>40</td>
<td>Saudi Arabia</td>
<td>1.28</td>
</tr>
</tbody>
</table>

China’s unweighted centrality ranks 7th and weighted centrality ranks 5th. China’s position in the collaboration network rises under the weighted centrality measure because China’s main collaborators are major countries (shown in Table 3) and our formula (2) gives more weight to the collaboration between China and those countries.

Like the unweighted case, Table 5 shows the increases of weighted centralities for the 20 countries in Table 2 between 2006-2010 and 2011-2015. As in Table 2, China has the largest centrality increase further demonstrating the rapid growth of China’s role and importance in the international collaboration network. In Table 2 the US’s centrality increase is nominal; however, based on weighted centrality, the US’s increase is significant and second only to China. This suggests that despite its already dominant position, in the past several years the US has continued to focus on increasing collaboration with other major countries. Curiously, Japan is the only country with flat or negative (weighted and unweighted) centrality growth. One reasonable interpretation of this is that China is eclipsing Japan as the leading Asian country in the international scientific collaboration network in recent years. We speculate that this may be a result of
China’s purposeful focus on growing its international influence and impact in science and innovation in hopes of driving further economic development, however further research is needed to validate this assertion.

Table 5. Increase of weighted centralities of 20 countries

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>USA</td>
<td>29.39</td>
<td>33.33</td>
<td>3.94</td>
</tr>
<tr>
<td>2</td>
<td>UK</td>
<td>15.71</td>
<td>17.89</td>
<td>2.18</td>
</tr>
<tr>
<td>3</td>
<td>Germany</td>
<td>15.34</td>
<td>16.59</td>
<td>1.25</td>
</tr>
<tr>
<td>4</td>
<td>France</td>
<td>11.26</td>
<td>12.23</td>
<td>0.97</td>
</tr>
<tr>
<td>5</td>
<td>Italy</td>
<td>8.44</td>
<td>10.09</td>
<td>1.65</td>
</tr>
<tr>
<td>6</td>
<td>Spain</td>
<td>6.50</td>
<td>8.79</td>
<td>2.29</td>
</tr>
<tr>
<td>7</td>
<td>China</td>
<td>7.45</td>
<td>11.98</td>
<td>4.53</td>
</tr>
<tr>
<td>8</td>
<td>Canada</td>
<td>9.00</td>
<td>9.60</td>
<td>0.60</td>
</tr>
<tr>
<td>9</td>
<td>Netherlands</td>
<td>6.24</td>
<td>7.78</td>
<td>1.54</td>
</tr>
<tr>
<td>10</td>
<td>Australia</td>
<td>5.71</td>
<td>8.27</td>
<td>2.56</td>
</tr>
<tr>
<td>11</td>
<td>Switzerland</td>
<td>5.56</td>
<td>6.92</td>
<td>1.36</td>
</tr>
<tr>
<td>12</td>
<td>Sweden</td>
<td>4.46</td>
<td>5.46</td>
<td>1.00</td>
</tr>
<tr>
<td>13</td>
<td>Japan</td>
<td>7.04</td>
<td>6.78</td>
<td>-0.26</td>
</tr>
<tr>
<td>14</td>
<td>Belgium</td>
<td>3.81</td>
<td>4.71</td>
<td>0.90</td>
</tr>
<tr>
<td>15</td>
<td>Russia</td>
<td>3.65</td>
<td>3.94</td>
<td>0.29</td>
</tr>
<tr>
<td>16</td>
<td>Poland</td>
<td>2.69</td>
<td>3.70</td>
<td>1.01</td>
</tr>
<tr>
<td>17</td>
<td>Brazil</td>
<td>2.62</td>
<td>4.07</td>
<td>1.45</td>
</tr>
<tr>
<td>18</td>
<td>Austria</td>
<td>2.69</td>
<td>3.83</td>
<td>1.14</td>
</tr>
<tr>
<td>19</td>
<td>Denmark</td>
<td>2.65</td>
<td>3.84</td>
<td>1.19</td>
</tr>
<tr>
<td>20</td>
<td>India</td>
<td>2.69</td>
<td>3.80</td>
<td>1.11</td>
</tr>
</tbody>
</table>

To better visualize the collaboration network, we choose to plot countries with large volumes of publications with international co-authorship. In figures 1 and 2, only countries with more than 30,000 such collaborations and connections with at least 20,000 collaborations in each of the five-year periods are plotted. In other words, the edges (collaboration) between two nodes (countries) are plotted only if the two countries have at least 20,000 collaborations. In Figures 1 and 2, the size of each node is proportional to the total number of co-authored publications and the width of the connection is proportional to the number of co-authored publications between the pair of countries.
In Figure 1, the US is clearly at the center of the network and the first primary collaborator of all countries except Switzerland in the five-year period 2006-2010. The top 3 collaborators of the US, measured by volume of publications, are the UK, Germany, and Canada. China’s total number of international collaborations is significantly lower than that of the US, the UK, Germany, and France between 2006 and 2010.

![Collaboration network between 2006 and 2010](image)

**Figure 1. Collaboration network between 2006 and 2010**

A very different view emerges when we compare Figure 2 with Figure 1. While the US is still the center of the network, the UK, Germany, France, Italy, Spain, and China all grow into sub-centers. China’s international collaboration volume increases so that it is close to that of the UK and Germany. The US and China are the only two countries with more than 20,000 co-authored papers with the UK, Japan, and Australia. This view clearly shows China’s emerging importance as a significant collaborator with other leading countries.
Conclusion and Discussion
Along with its bustling economy, China has rapidly increased its productivity of academic publication output over the past decade to the point where its annual volume of papers is second only to the United States; and the gap is quickly narrowing. At the same time, global collaboration (as measured through papers with international co-authorship) in scientific research has intensified. Researchers in scientometrics and research policy, the Chinese government, and the public are all interested in China’s growing prominence in the international scientific collaboration network accompanied by the broader economic implications of this emergence. This paper endeavored to quantify this phenomenon using novel bibliometric measures.

Commonly used measures of a node’s role and importance in a network, such as degree centrality, closeness centrality, and betweenness centrality are based in general social network analysis. However, different from a general social network, the scientific collaboration network of countries is a complete graph where any two countries have connections (collaboration relationships.) Thus, traditional measures of nodes cannot be realizably applied to analyze the scientific collaboration network of countries.

We proposed two novel indicators to measure the role of countries in the global collaboration network. In formulating these two indicators, we presumed that if the collaborations between countries $i$ and $j$ account for a large proportion of $j$’s total collaborations, then $i$ plays an important role among all $j$’s collaborators. Although these measure help to quantify China’s emergence in importance in the network, they still do not show China’s collaborative role in line with its position as second (only behind the
US) in overall publication output. China’s increase in importance in the past decade is the most significant among the 20 countries with the highest overall publication output. China is clearly emerging as a sub-center for substantial international scientific collaboration.

It is well established in the existing literature that international scientific collaboration is complex and affected by geographic, linguistic, cultural, sociological, and economic factors. While we revealed the roles of several leading countries in an aggregate quantitative way, we did not analyze any of the myriad underlying factors that very likely impact the actual collaborations. There is certainly opportunity for further deep study in these areas for interested researchers. In addition, we analyzed international collaborations based on publications across a broad range of academic disciplines. It is also probably worth studying collaborations in individual fields for a more nuanced understanding of the distinctions within discipline-specific sub-networks; we imagine that these can differ greatly and such study could reveal further interesting and instructive insights.

In this paper we only consider and measure the collaboration relationship from the perspective of bilateral relations. For the collaborations between A and B, although many other countries also participated in their coauthored papers, we disregard such multilateral collaboration information when we consider the collaborations between A and B. To describe the complete collaboration relationships by taking the full coauthorship information in the papers into account, we need more complex methods, indicators, and multilateral collaboration visualization. This is an important topic and it’s worth deep exploring in the future.

References


Wilson, R. J. (2012). *Introduction to Graph Theory* (5th ed.): Pearson Publishing Ltd.


Interdisciplinarity and collaboration: On the relationship between disciplinary diversity in the references and in the departmental affiliations

Lin Zhang¹,²*, Beibei Sun¹, Ying Huang³, Lixin Chen¹

¹Zhanglin_1117@126.com
¹Dept. Management and Economics, North China University of Water Resources and Electric Power, China
²Centre for R&D Monitoring (ECOOM) and Dept. MSI, KU Leuven, Belgium
³School of Management and Economics, Beijing Institute of Technology, Beijing, China.

Abstract: We proposed a novel conceptual framework to investigate two different aspects of interdisciplinarity in the process of knowledge integration: the subject categories referenced and the specialist disciplines of the authors’ listed affiliations. Diversity was used to capture the disciplinary heterogeneity of our data sample in both respects. We presented an explorative methodology for retrieving feature words to classify the affiliates disciplines. Some interesting observations were drawn from this pilot study. 1) In general, there are positive relationships between interdisciplinarity and collaboration, but the specific nature of those relationships varies among different journals and according to different indicators. 2) Based on a schematic model of the discipline diversity in references versus the discipline diversity in affiliations, the combinations of ‘low-low’ and ‘high-high’ occur much more frequently than ‘low-high’ or ‘high-low’ in each of the journals under study.

Conference topic
Methods and techniques; Indicators; Interdisciplinarity

Introduction

Interdisciplinary research (IDR) is generally seen as a source of creativity and innovation (Dogan & Pahre, 1990). And, there is a general consensus that what differentiates IDR from multidisciplinary research is the degree of knowledge integration (Porter et al. 2008). As such, appropriately identifying interdisciplinary activity is an important step in understanding where it occurs and that it is properly assessed (Rousseau et al. 2017). According to the National Academy of Science in the US (2004), IDR is:

“a mode of research by teams or individuals that integrates information, data, techniques, tools, perspectives, concepts and/or theories from two or more disciplines or bodies of specialized knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or area of research practice”.

The above definition emphasizes knowledge integration on the part of individuals, teams, departments, institutions, regions, and even countries. The more an article, or any other investigative piece, integrates different sources, the more interdisciplinary it becomes. According to Grant et al. (2015), two of several ingredients that generate excellent outcomes in IDR are:

- Diversity in research fields: a broad range of disciplines supports exceptional levels of research excellence.
- Diversity of research organizations: where mission-led units complement large and small universities with regional as well as international engagement.

Various bibliometrics indicators have focused on measuring interdisciplinarity in terms of the knowledge that originates from different research fields. Categorical analysis has predominantly relied upon lists of cited references (e.g., Rafols & Meyer, 2010; Wang et al., 2015, Zhang et al., 2016). However, measuring the diversity of research organizations has not been widely explored.
In a recent study, Adams et al. (2016) noted that the same project may be indexed as interdisciplinary according to one parameter, say the diversity of its references, but not for another, such as its authors’ affiliations or the universities the study originated from. It is not surprising that different indicators can deliver inconsistent, or even contradictory, results. For this reason, Adams et al. (2016) points out that it is essential to consider a framework for analysis that draws on multiple indicators, rather than expecting any simplistic index to produce an informative outcome on its own.

The research questions in this study arise from Grant et al.’s (2015) IDR progenitors – diversity in research fields and diversity in research organizations. We not only measure the interdisciplinarity of research studies through their cited references, but we also try to identify diversity in the listed affiliations of the studies’ co-authors at a departmental level. Our two main research questions are:

1. Can IDR be achieved by single authors or one research organisation, or does IDR normally require collaboration between different entities?

2. What is the relationship between disciplinary diversity in the references of a study and the affiliations of the study’s co-authors?

Data and Methods

Data

For the purposes of this study, all the article-style publications in three journals – Scientometrics, Bioinformatics and the Journal of the American Geriatrics Society – were downloaded from Clarivate Analytics’ Web of Science (2006-2015). These three journals were chosen following the strategy in our previous study (Zhang et al., 2016). In total, our sample dataset contained 11,883 records. The reference for each paper were retrieved using a simple search from Web of Science. The authors’ affiliations were obtained by first identifying feature words in the affiliates’ names from full attributions of the authors. The feature words were: Dept, Sch, Ctr, Coll, Inst, Lab, Assoc, Fac, Div, etc. Then, the thesaurus function in Thomson Data Analyzer’s (TDA) text mining tool was used to clean the information. Because our research question focuses on affiliations at the departmental level, but some of the feature words are also common to university and institution names, e.g., the Harbin Institute of Technology, a thorough manual check was conducted to reduce the effect of noise. After discarding the records that did not include affiliations or lacked relevant feature words, 9572 affiliations were confirmed from the bylines of 11,318 articles.

Methods

Setting the Discipline classification systems

It is clear that using the term IDR, first requires a definition of the term discipline. Here, we use the term discipline as a synonym for a research specialty or field of study. There are already several pre-defined ‘intellectual’ classification schemes in wide use in bibliometrics, such as the 22 broad field classifications used in the Essential Science Indicators database or the 250+ subject categories used by Web of Science (WoS). Glänzel and Schubert (2003) proposed, and subsequently developed, a hierarchical scheme: the Leuven–Budapest (ECOOM) subject classification. We chose to adopt both the WoS and ECOOM classifications as our pre-defined benchmarking systems. ECOOM’s hierarchical structure makes it meaningful for building a matching and comparison system, while the 250+ categories in WoS can all be assigned to one of 68 subfields or 16 major fields in ECOOM.
The WoS category corresponding to the references in the articles was obtained directly from the WoS database, and no further classification or reclassification processes were applied, beyond reducing the WoS categories to a set of 245 relevant subject areas, as explained in the affiliate classification process.

The most challenging task was to match the affiliations with the discipline most relevant to their work. After retrieving all the subject-related feature words from the 9572 affiliate names through several rounds of manual and machine-based processing, we firstly tried to match each affiliate to one of ECOOM’s 16 major fields. ECOOM was chosen over the WoS classifications because the subject-related feature words in the affiliate names were not specific enough to allow proper matches in many cases. However, some adjustments to ECOOM’s system were required to completely match all the records, as shown in Table 1.

Table 1. The relationships between ECOOM’s major disciplines and our corresponding major disciplines used to classify the affiliations

<table>
<thead>
<tr>
<th>Areas</th>
<th>ECOOM’s major disciplines</th>
<th>The corresponding major disciplines used to classify affiliates (DA_ fields)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multidisciplinary Sciences</td>
<td>Computer Science &amp; Information Technology</td>
<td></td>
</tr>
<tr>
<td>Natural Sciences</td>
<td>Agriculture &amp; Environment</td>
<td>Agriculture &amp; Environment</td>
</tr>
<tr>
<td></td>
<td>Biology (Organismic &amp; Supraorganismic Level)</td>
<td>Biology</td>
</tr>
<tr>
<td></td>
<td>Biosciences (General, Cellular &amp; Subcellular Biology; Genetics)</td>
<td>Biology</td>
</tr>
<tr>
<td></td>
<td>Biomedical Research</td>
<td>Biology Medicine</td>
</tr>
<tr>
<td></td>
<td>Chemistry</td>
<td>Chemistry</td>
</tr>
<tr>
<td></td>
<td>Clinical &amp; Experimental Medicine I (General &amp; Internal Medicine)</td>
<td>Medicine</td>
</tr>
<tr>
<td></td>
<td>Clinical &amp; Experimental Medicine II (Non-Internal Medicine Specialties)</td>
<td>Medicine</td>
</tr>
<tr>
<td></td>
<td>Engineering</td>
<td>Engineering</td>
</tr>
<tr>
<td></td>
<td>Geosciences &amp; Space Sciences</td>
<td>Geosciences &amp; Space Sciences</td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>Mathematics</td>
</tr>
<tr>
<td></td>
<td>Neuroscience &amp; Behavior</td>
<td>Medicine; Psychology</td>
</tr>
<tr>
<td></td>
<td>Physics</td>
<td>Physics</td>
</tr>
<tr>
<td>Social Sciences and Humanities</td>
<td>Arts &amp; Humanities</td>
<td>Arts &amp; Humanities</td>
</tr>
<tr>
<td></td>
<td>Social Sciences I (General, Regional &amp; Community Issues)</td>
<td>Social Science, General</td>
</tr>
<tr>
<td></td>
<td>Social Sciences II (Economics, Politics &amp; Legal Sciences)</td>
<td>Social Science (Economics &amp; Management)</td>
</tr>
</tbody>
</table>

As a result, ECOOM’s 16 major fields were reduced to a total of 13 major disciplines: 1) Computer Science & Information Technology, 2) Agriculture & Environment, 3) Biology, 4) Medicine, 5) Chemistry, 6) Engineering, 7) Geosciences & Space Sciences, 8) Mathematics, 9) Psychology, 10) Physics, 11) Arts & Humanities, 12) Social Sciences, General, 13) Social Sciences (Economics & Management).

Some brief notes about the adjustments made to ECOOM’s 16 major disciplines follow:

1) Multidisciplinary Sciences was excluded. Given our main research question is to study interdisciplinarity in research, each departmental affiliation needed to be classified into a
specific discipline. Therefore, the field of ‘Multidisciplinary Sciences’ is not applicable to this study.

2) The two medicine fields were merged into a single discipline, Medicine. The two fields of biology and bioscience were merged into a single discipline, Biology. This change was made because many departmental affiliations contain broad subject-related feature words, such as ‘Dept Med’ and ‘Dept Bio’. Hence, Biology and Medicine were treated as two independent disciplines in our classification scheme.

3) Some interdisciplinary fields in ECOOM were divided into specific disciplines. Biomedical Research was divided into Biology and Medicine; Neuroscience & Behaviour was divided into Medicine and Psychology. Note that in cases where an affiliate’s name contained a word related to biomedical, it was assigned to both Biology and Medicine in the new system. In other words, multi-assignment was allowed.

4) Psychology was added as an independent discipline. In recent years, there has been a great deal of development in psychological research, and the trends within this field are commonly found in the names of affiliates all over the world.

5) Computer Science & Information Technology was added as a new major discipline. Our data sample included many studies on computer science and information technology, and related feature words were frequently found in the affiliate names. To adapt the data distribution in our sample to maintain a relative balance between each discipline, we separated Computer Science & Information Technology from Engineering and Social Sciences I (General, Regional & Community Issues).

6) Social Science I & II were divided into two specific disciplines: Economy & Management and General Social Sciences. This adjustment was also made to even out the data distribution of the affiliate feature words within our sample.

As previously mentioned, every major field and sub-field in ECOOM has a strong match to the categories in WoS. WoS tags 252 categories to 12,000+ journals in its 2016 edition. The six newest categories, ‘Audiology & Speech-language Pathology,’ ‘Cell & Tissue Engineering,’ ‘Ergonomics,’ ‘Green & Sustainable Science & Technology,’ ‘Logic,’ and ‘Nanoscience & Nanotechnology’, were only added in 2016 and are not included in ECOOM’s classification scheme. These categories were removed, along with multidisciplinary sciences as mentioned above, to arrive at a final list of 245 WoS categories. These were synchronized with the 16 major ECOOM fields to further help classify the affiliates into our 13 disciplines.

Assigning departmental affiliations to their corresponding discipline(s)

Discipline-specific lexicons for the 245 WoS categories were used to correlate the discipline feature words retrieved from the affiliate names with a field of research. All feature words were carefully checked using a combination of manual and machine-based work to build a vocabulary of feature words for each major discipline, and this list was used to assign the affiliates to one or more of the 13 disciplines in our classification system.

Throughout the entire process, we focused on a balance between data precision and the recall rate, and this trade-off between accuracy and efficiency was the primary motivator for the decision to use feature words as our vocabulary. After several rounds of manual and machine-based screening, 379 discipline feature words were collected from the affiliates’ names and 8444 departmental affiliations were identified and classified into the 13 different disciplines. Records without listed
affiliations were excluded, resulting in 10,899 articles for further analysis, i.e., 91.64% of the original data sampled.

We attempted to obey some rules in our processing procedure:

1) Unknown words in affiliate names were classified using official sources. Generally, the abbreviations of academic terms in affiliation names, such as Med, Bio, Stat, made it easy to identify their discipline(s). However, some words, e.g., ‘Fis’ and ‘Quim’, were harder to classify. In these cases, a manual trace of the author’s full affiliations in the original document to the affiliate’s official website was conducted to confirm their field of research manually. In the example above, ‘Fis’ and ‘Quim’ were identified as the Spanish abbreviations for ‘Física’ and ‘Química’, and the corresponding English words are Physics and Chemistry.

2) ‘Full matching’ technology was applied whenever needed. For instance, the abbreviation term ‘Dis’ is an abbreviation for the medical feature word ‘Disease’. However, ‘Dis’ may also appear in ‘ADIS’ or ‘Discovery’, etc. The same applies to other abbreviations, such as ‘eye’, ‘soc’, ‘eth’, ‘art’, ‘gene’. In these cases, ‘full match’ was set for these particular words to avoid possible noise.

3) Where an abbreviation may have two meanings, the full affiliate name was used. For example, when the feature word ‘Vet’ appears alone in an affiliate name, it usually means Veterinary and should be classified into Biology. However, when it appears within the phrase ‘Vet Affairs’ or ‘Vet Admin’, it represents ‘Veterans’, and the related affiliate was generally found to be a veterans rehabilitation center, which was classified into Medicine. A manual check of the full affiliate name resolved each case.

4) Affiliates with more than one discipline feature word in their name, like ‘Dept Biochem’ and ‘Div Biomed Stat & Informat’, were regarded as interdisciplinary affiliations and were given multiple assignments into each corresponding discipline. For example, ‘Dept Biochem’ was assigned to both ‘Biology’ and ‘Chemistry’.

**Applying interdisciplinarity measure indicators**

Among the various interdisciplinarity measures available, we focused on measuring diversity through its three components: variety, balance, and disparity (Stirling 2007). These three decompositions make it possible to explore different components of diversity in the cited references of a corpus of publications and the affiliations of the authors. We selected the true diversity measure (TD) proposed in Zhang et al. (2016) as the representative indicator of integrated diversity, and the number of different subject categories as the decomposed measurement representing variety in the concept of diversity.

**TD** is defined as

$$TD = \frac{1}{\sum_i p_i p_j (1-d_{ij})}$$  \hspace{1cm} \text{(1)}

where $i$ is the index, $x_i$ is the number of references/affiliations to the $i$-th subject category, and the subject categories are sorted by $x_i$ in non-decreasing order. $p_i = x_i / \sum x_i$. $d_{ij}$ is the dissimilarity between the subject categories $i$ and $j$. $d_{ij} = 1 - s_{ij}$, where $s_{ij}$ is the Salton’s cosine similarity between the subject categories $i$ and $j$ based on the WoS subject cross-citation similarity matrix (1991-2015). When the 13 disciplines in the departmental affiliations are used, the WoS subject cross-citation matrix is aggregated into the ‘DA_ fields’ (see in Table 1) similarity matrix. $d_{ii} = 0$ for all $i$. Detailed information and supplementary materials can be found at http://suo.im/1o7Ff9 or
Figure 1 presents an outline of the data processing methods used in this study.

The following abbreviations are used for the four diversity measures applied in this study:

- $N_{SF\_A}$ denotes the “number of different discipline fields in the affiliations”;
- $N_{SC\_R}$ denotes the “number of different subject categories in the cited references”;
- $TD_{A}$ represents the “true diversity measure of the affiliations”;
- $TD_{R}$ represents the “true diversity measure of the references”;

Affiliate(s) always refers to the departmental affiliation(s) listed by an author.

Also, note that WoS’s 245 categories were used to classify the references, whereas the affiliates were classified into the 13 major disciplines derived and adjusted from ECOOM’s classifications.

**Results**

**Descriptive statistics**

Through a count of the co-authors, listed affiliations, and countries in the attributions for each paper in the three selected journals, and based on the data processing and the classification systems presented in the Methods section, we assigned all the affiliations and references into their corresponding disciplines and calculated the $N_{SF\_A}, N_{SC\_R}, TD_{A}, TD_{R}$ for each individual
paper. The descriptive statistics for each of the measures in the three journals is shown in Table 2. It’s interesting to see that the Journal of the American Geriatrics Society (JAGS) had the largest collaboration scale in terms of both the number of authors and the average affiliations, while Bioinformatics had the largest average values across all the diversity measures (N_SF_A, N_SC_R, TD_A, TD_R). This difference was due to the fact that Bioinformatics has high interdisciplinarity in both affiliations and cited references, and as a contrast, the large collaboration in JAGS are more based on same or similar disciplines. Scientometrics had a lower level of both author and affiliate collaborations.

Table 2. Descriptive statistics for the different measures in the three selected journals

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of authors</td>
<td>mean 4.47</td>
<td>2.73</td>
<td>6.15</td>
</tr>
<tr>
<td></td>
<td>median 4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Number of affiliations</td>
<td>mean 2.12</td>
<td>1.86</td>
<td>3.28</td>
</tr>
<tr>
<td></td>
<td>median 2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Number of countries</td>
<td>mean 1.36</td>
<td>1.33</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td>median 1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>N_SF_A / N_SC_R</td>
<td>mean 2.58/11.04</td>
<td>1.849/8.763</td>
<td>2.141/9.416</td>
</tr>
<tr>
<td></td>
<td>median 2/11</td>
<td>2/8</td>
<td>2/9</td>
</tr>
<tr>
<td>TD_A / TD_R</td>
<td>mean 1.848/6.402</td>
<td>1.599/4.73</td>
<td>1.357/4.845</td>
</tr>
<tr>
<td></td>
<td>median 1.87/6.455</td>
<td>1.671/4.217</td>
<td>1.26/4.617</td>
</tr>
</tbody>
</table>

The number of affiliations and countries is highly skewed, so we adopted a dummy variable. If the attributions in a paper included more than one affiliation or country, the variable was set to 1, and 0 otherwise. The same applied if a paper was written by multiple authors or only one author. Table 3 presents the collaboration ratios for each journal. The author collaboration ratio represents the percentage of multi-authored papers. The cross-affiliate collaboration ratio reflects the percentage of papers written by co-authors with different affiliations. The international collaboration ratio shows the percentage of papers written by co-authors from different countries. And, the cross-discipline affiliation ratio shows the percentage of papers written by co-authors with affiliations across different disciplines.

It is interesting to note that the JAGS had both the highest author collaboration ratio (98.11%) and cross-affiliation collaboration ratio (80.29%), but the lowest international collaboration ratio (16.01%). This suggests that papers in the JAGS are highly collaborative, but these collaborations are mostly on a national level. Bioinformatics had the highest international collaboration ratio (27.62%), and cross-discipline affiliations ratio (79.10%), which indicates that the articles in Bioinformatics have a higher international and interdisciplinary orientation in terms of collaboration. The co-authorship, cross-affiliation, and cross-discipline ratios for Scientometrics were the lowest of the three journals.

Table 3. The collaboration ratios for authors, affiliations, countries, and disciplines

<table>
<thead>
<tr>
<th></th>
<th>Author collaborations</th>
<th>Cross-affiliation collaborations</th>
<th>International collaborations</th>
<th>Cross-discipline affiliations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioinformatics</td>
<td>97.53%</td>
<td>59.41%</td>
<td>27.62%</td>
<td>79.10%</td>
</tr>
<tr>
<td>Scientometrics</td>
<td>80.13%</td>
<td>55.58%</td>
<td>25.84%</td>
<td>55.86%</td>
</tr>
<tr>
<td>J. Am. Geriatr. Soc.</td>
<td>98.11%</td>
<td>80.29%</td>
<td>16.01%</td>
<td>61.39%</td>
</tr>
<tr>
<td>All</td>
<td>94.71%</td>
<td>63.96%</td>
<td>24.45%</td>
<td>70.88%</td>
</tr>
</tbody>
</table>
Figure 2 demonstrates the distribution of individual articles versus the number of affiliations, the number of different disciplines among the affiliates, and the number of different subject categories in the cited references. In *Bioinformatics* and *Scientometrics*, articles with a single affiliation dominate. By contrast, the *JAGS* shows a rather different distribution, where cross-affiliate collaboration is common. However, it is interesting to note that most collaborations are within a single discipline. The level of cross-discipline affiliations in *Bioinformatics* is also remarkable, given their low proportion of cross-affiliations. Single-discipline affiliations are dominant in both the *JAGS* and *Scientometrics*. In terms of the variety of disciplines in the cited references, *Scientometrics* appears to have the lowest interdisciplinarity compared to the other two journals. However, the distribution of $N_{SC\text{-}R}$ in *Scientometrics* has a rather long tail to the right, and the article with most reference subjects – a total of 64 categories – was found in this journal.

Fig.2 The distribution of individual papers according to the number of affiliations, $N_{SF\text{-}A}$ and $N_{SC\text{-}R}$ in the three journals under study
Is IDR more achieved by collaboration among different entities?
To answer this research question, we divided all the articles in each journal into groups according
to the number of authors, affiliations, countries, and affiliate disciplines, then calculated
the average number of reference subjects for each group. Figure 3 shows the results of the comparison
between the three different journals. Both Bioinformatics and the JAGS show a general increase in
interdisciplinarity in line with an increasing number of authors, affiliations, affiliate disciplines,
and the number of referenced subjects. In Bioinformatics’ case, these rising trends are rather
striking. Against expectations, no specific relationship was found between reference diversity and
the collaboration ratios for Scientometrics. Since the number of countries is highly skewed, we
separated the articles published within a single country from those with international collaborations.
In general, the papers in Bioinformatics and the JAGS with international collaborations contained
a more diverse set of reference subjects but Scientometrics, again, presented an exceptional case.

We further calculated the average TD_R for each group according to the number of authors and
affiliate disciplines. Figure 4 presents the results. Similar to our observations in Figure 3, the JAGS
and Bioinformatics both showed a rising trend in TD_R, along with an increase in the number of
co-authors, affiliate disciplines – and, again, exceptionally so for Bioinformatics. Yet, once more,
Scientometrics presents a different profile. Combined with the previous analysis, Scientometrics
generally shows a lower level of collaboration than the other two journals, and the relationships
between collaborations and interdisciplinarity are rather inexplicit.

Fig. 3 The relationships between
the number of referenced subjects and the different collaboration measures
Fig. 4 The relationship between TD_R and the different collaboration measures in the three journals

Disciplinary diversity in references vs. disciplinary diversity in affiliations

In this section, we explore knowledge integration from two different perspectives – disciplinary diversity in references versus disciplinary diversity in affiliations. Fig. 5 provides a schematic representation of this twofold perspective, following Porter et al. (2007) and Rafols and Meyer’s (2010) proposals. Since this study examines individual publications and knowledge integration through the disciplines associated with author affiliations and references, the central node in each graph represents the focal literature. The affiliate discipline sets are on the left, and the reference discipline sets are on the right. Each of the colored icons in the network represents a discipline. There are four possible combinations:

(i) low affiliate discipline diversity – low reference discipline diversity in the references (LDA-LDR) represents cases of specialized disciplinary research, where all references are from same or similar disciplines, accomplished by authors from the same or similar affiliation disciplines.

(ii) low affiliate discipline diversity – high reference discipline diversity (LDA-HDR) indicates cases where the references cite a variety of subjects, but the authors are affiliated with the same or similar disciplines.

(iii) high affiliate discipline diversity – low reference discipline diversity (HDA-LDR) represents cases of specialized research, where all references cite the same or similar disciplines, but the authors are affiliated with a variety of disciplines.

(iv) high affiliate discipline diversity – high reference discipline diversity (HDA-HDR) denotes cases where the references cover a variety of subjects, accomplished by authors with a variety of affiliation disciplines.

Tables 4 and 5 show that the ‘low-low’ and ‘high-high’ combinations occur much more frequently than ‘low-high’ or ‘high-low’ in each of the journals under study. It is interesting that Scientometrics follows a similar pattern to the other two journals across all quadrants, even though the relationship between the interdisciplinarity and collaboration measures of their papers was obscure in our previous analyses. Bioinformatics presents the only somewhat exceptional case. When using the integrated diversity measure (TD), the distribution among the four combinations is relatively even. This observation deserves deeper exploration in the future.

To illustrate the schematic model using our empirical data, we ranked all the individual publications in each journal according to the two measures of variety (N_SF_A, N_SC_R) and the two measures of integrated diversity (TD_A, TD_R) in descending order. The top third of all publications for each measure were regarded as “high,” and the bottom third were regarded as
“low”. After a cross-matching process between the high and low groups based on each measure, we derived the representative paper samples for each combination as shown in Fig. 5. The proportion of articles for each of the four combinations are shown in Tables 4 and 5.

<table>
<thead>
<tr>
<th>Disciplinary diversity in references</th>
<th>low</th>
<th>high</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disciplinary diversity in affiliations</strong></td>
<td>Specialized disciplinary research with interdisciplinary affiliations</td>
<td>Highly interdisciplinary research</td>
</tr>
<tr>
<td>high</td>
<td>![Diagram of high interdisciplinary affiliations]</td>
<td>![Diagram of highly interdisciplinary affiliations]</td>
</tr>
<tr>
<td>low</td>
<td>![Diagram of low interdisciplinary affiliations]</td>
<td>![Diagram of specialized interdisciplinary affiliations]</td>
</tr>
</tbody>
</table>

Fig.5 Disciplinary diversity in references vs. Disciplinary diversity in the affiliations

Table 4. The proportion of publications in each combination based on N_SF_A and N_SC_R

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA-LDR</td>
<td>19.94%</td>
<td>14.12%</td>
<td>17.53%</td>
</tr>
<tr>
<td>LDA-HDR</td>
<td>5.63%</td>
<td>3.96%</td>
<td>3.00%</td>
</tr>
<tr>
<td>HDA-LDR</td>
<td>4.90%</td>
<td>6.64%</td>
<td>5.44%</td>
</tr>
<tr>
<td>HDA-HDR</td>
<td>18.29%</td>
<td>18.53%</td>
<td>16.31%</td>
</tr>
</tbody>
</table>

Table 5. The proportion of publications in each combination based on TD_A and TD_R

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA-LDR</td>
<td>12.15%</td>
<td>13.67%</td>
<td>16.68%</td>
</tr>
<tr>
<td>LDA-HDR</td>
<td>10.18%</td>
<td>3.74%</td>
<td>4.25%</td>
</tr>
<tr>
<td>HDA-LDR</td>
<td>10.17%</td>
<td>9.65%</td>
<td>7.58%</td>
</tr>
<tr>
<td>HDA-HDR</td>
<td>12.25%</td>
<td>12.61%</td>
<td>14.05%</td>
</tr>
</tbody>
</table>

Conclusions

In this article, we proposed a novel conceptual framework to investigate two different aspects of interdisciplinarity in the process of knowledge integration: the subject categories referenced and the specialist disciplines of the authors’ listed affiliations. Diversity was used to capture the disciplinary heterogeneity of our data sample in both respects. We presented an explorative methodology for retrieving feature words to classify the affiliates disciplines, based on a
Some interesting observations were drawn from this pilot study. 1) In general, there are positive relationships between interdisciplinarity and collaboration, but the specific nature of those relationships varies among different journals and according to different indicators. For example, interdisciplinarity in the articles in the Journal of the American Geriatrics Society and Bioinformatics is increasing, in line with an increase in the number of authors, affiliations, and affiliate disciplines, but Bioinformatics is showing exceptional interdisciplinary growth. Whereas, the relationship between collaboration and interdisciplinarity for Scientometrics is not at all explicit.

2) Based on a schematic model of the discipline diversity in references versus the discipline diversity in affiliations, the combinations of ‘low-low’ and ‘high-high’ occur much more frequently than ‘low-high’ or ‘high-low’ in each of the journals under study.

We feel this research is beneficial to the study of interdisciplinarity. Our future research will adapt the proposed approaches for use with larger bibliometric sets.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (Grant No.71573085) and the Innovation Talents of Science and Technology in HeNan Province (Grant No.16HASTIT038).

References


Abstract

At early 2016 the new index was launched on Web of Science platform — Russian Science Citation Index (RSCI). The database is free for all Web of Science subscribers except those from the former Soviet Union countries. This database includes publications from 652 best Russian journals and is based on the data from Russian national citation index — Russian Index of Science Citation (RISC). RISC was launched in 2005 but there is very limited information about it available in English-language scholarly literature by now. The aim of this paper is to describe the history, actual structure and user possibilities of RISC. We focus on the novel features of RISC which are crucial to bibliometrics and unavailable in international citation indices.

Introduction. History and main objectives of RISC

RISC was launched in 2005 as a government-funded project primarily aimed at creating a comprehensive bibliographic/citation database of Russian scholarly publishing for evaluation purposes based on Scientific Electronic Library (further eLIBRARY.RU) which started as a full-text database of scholarly literature for grant holders of Russian Basic Research Foundation. The main purpose of eLIBRARY.RU was to provide Russian scientists with papers from leading academic journals (Eremenko, 2003).

The evaluation of research implies the bibliometric analysis of research output of scientists, organizations, etc. At the same time only less than 10% of publications of Russian scientists are included into international citation indexes, i.e. Web of Science and Scopus (Tretyakova, 2015). Almost all publications in social sciences and humanities are in Russian — and they are absent from international databases. The primary aim of the federal target scientific-technical program was to collect and index information from all available publications in Russian scholarly journals and their reference lists to make possible the further evaluation of journals’ quality and research output of Russian scientists, universities and research organizations.

In general, motives for creating national citation indexes are explained by Pislyakov (2007). There are plenty of examples and stories of implementation of such databases for different countries. For example, in China even several national indexes were created (Jin & Wang, 1999; Su, Deng & Shen, 2014; Wu et al., 2004; Ye, 2014). There are also cases in India (Yadav & Yadav, 2014), Japan (Negishi, Sun & Shigi, 2004) etc. Other stories were not so successful, the Serbian case was presented at previous ISSI conferences (Šipka, 2005; Pajić, 2015).
Unlike Web of Science and Scopus there are no strict criteria for journals to be indexed in RISC. At the very beginning of the project it was decided to index any scholarly journal. Nevertheless, the principle of indexing can not be called purely declarative: eLIBRARY.RU refuse indexing the propaganda journals or magazines that were deemed to be just popular or commercial media designed for a wide non-scientific audience.

It is important to differ the journals at the eLIBRARY.RU platform from journals indexed in RISC. Journal catalogue of eLIBRARY.RU now includes 12975 Russian periodicals, only 4847 of them being indexed in RISC. RISC indexes also 541 foreign journals which are published mainly in former Soviet countries.

The research output can not be evaluated only by journal publications — the scientific activity is also reflected in different book publications, conference materials, dissertations, etc. In 2013 the book project was launched in RISC and now eLIBRARY.RU platform hosts non-periodical publications: monographs, reference books and dictionaries, textbooks and manuals, collections of articles, conference proceedings, theses and dissertation abstracts.

The problem of author identification is crucial for all citation databases and there are different ways to solve it. RISC involves authors to resolve this problem with the service Science Index for authors since 2011 and for organizations since 2012 (Arefiev, Eremenko & Glukhov, 2012).

In 2015, the collection of the best journals of all journals indexed in the RISC was created for launching on the Web of Science platform in a separate database Russian Science Citation Index by agreement with the company Thomson Reuters IP (now — Clarivate Analytics). It has greatly expanded the presence of Russian scientific journals in the international information space, especially for periodicals in the field of social and humanitarian, technical and medical sciences, which were underrepresented in the Web of Science and Scopus. At the same time the RISC core has been selected. It includes the best publications that may be used to evaluate the effectiveness of scientific research based only on the highest quality segment of the scientific works of Russian scientists.

![Figure 1. The distribution of publication types in RISC](image)

- Journal article (71%, 17 mln)
- Dissertation (10%, 2.7 mln)
- Conference paper (9%, 2.2 mln)
- Book (4%, 9.7 mln)
- Patent (2%, 0.5 mln)
- Book chapter (2%, 0.5 mln)
- Conference Proceedings (<1%, 0.19 mln)
- Research report
- Preprint
Content and coverage. Information sources

There are 25 million documents at eLIBRARY.RU platform with almost 260 million references. The distribution of publication types is shown in Fig. 1.

eLIBRARY.RU contains almost 60,000 titles, 14,500 of them are Russian. Full texts are indexed for 4391 journals, this enables extracting the citation context (in the form of snippets), the unique feature of RISC unavailable in other citation databases. Each year the number of journals in eLIBRARY.RU is growing approximately 300 journals per year (Fig. 2).

In 2010 an agreement was reached with the leading international publisher of scientific literature Elsevier to import into RISC the information about Russian authors and to attribute to them their works from the international citation database Scopus. This makes possible a comprehensive evaluation of publication output and citation analysis of Russian scientists and scientific organizations. Since then RISC takes into account not only papers in Russian journals indexed in the database, but also Russian papers in international journals.

Books come to eLIBRARY.RU from several sources. The main are publishers’ or authors’ agreements with eLIBRARY.RU. There are 1,750 contracts with 1,500 Russian and foreign publishers to deposit books to eLIBRARY.RU platform and more than 5,000 contracts with the authors to include their monographs.

The main source for patent information is Federal Institute for Industrial Property, the research reports are mainly imported from database of Russian Foundation for Basic Research. The whole number of documents in eLIBRARY.RU and their distribution by document types is shown in Fig. 3.
Figure 3. Annual number of documents in eLIBRARY.RU since RISC launching

Structure

Publication metadata

Every publication in RISC is attributed to specific publication type: research article, review, short communication, conference paper, letter, research report, book review, annotation, letter, editorial, biographical item; monograph, article collection, textbook, dictionary or reference book, brochure, guidelines; conference proceedings; dissertations; patents; grant report; preprint.

Publication metadata may include different identifiers — DOI, UDC, PMID, etc. As usual, the abstract and keywords in Russian and English are provided by publishers together with the references.

The preview of each publication is supplied with its bibliometrics and altmetrics indicators. The registered users may also see the links for citations in Google scholar. The links for the full text at the external sites are also available if publisher provides this information.

The bibliometric indicators include number of citations in RISC, Web of Science and Scopus, the citedness normalized by journal and subject area, other essential information. Altmetrics show the number of previews, downloads at eLIBRARY.RU and the number of article sets created by users to which this paper was added.

Lists of references have links to the publications included into database and citation snippets (if their full text is available at eLIBRARY.RU). It is possible thanks to the important difference of RISC from Web of Science and Scopus: RISC is not only fully integrated with eLIBRARY.RU platform, but is a part of it. Bibliographic and bibliometric information are combined with the service of online electronic library, providing access to the full texts which are stored in the system together with metadata. Today almost 4,500 journals use eLIBRARY.RU as a full-text
hosting platform. Citation snippets is a unique feature of RISC comparing with Web of Science, Scopus or Google Scholar.

Author profiles
At the moment of registration the scientist inputs his/her personal data, affiliations, name variants, research interests, list of journals, where he has publications, ResearcherID, Scopus, ORCID identifiers. This helps to identify the authors in papers. Each registered author receives a SPIN-code — Scientist Personal Identification Number. The number of registered authors now exceeds 430 thousand.

For each author, we may find all documents at eLIBRARY.RU where the scientist is identified as an author, editor or reviewer, the list of citing publications and analysis of author’s publication output. The metrics are available: the number of papers and the citations received by them, h-index, bibliometric indicators for the last 5 years and distribution of publications by types, journals, coauthors, etc. Also there are the links to scientist profiles in external systems (Scopus, ResearcherID, ORCID, etc).

The publications extracted from reference lists in the full author publication list is the second unique feature of RISC in comparison with Web of Science and Scopus, where such publications can be found only by special option of reference search as secondary documents. This option is also realized in Google Scholar.

Organization profile
Organization profile includes the general official information about organization and its publication output.

Publication data include lists of documents with organizational affiliation, all publications of authors during the period of their work in organization and all publications of current staff regardless of affiliation mentioned. These options help to evaluate research output in organization and potential opportunities.

Analysis of publication activity of organization overviews the whole number of publications in RISC and citations received, number of authors affiliated with organization and other characteristics. Also, one can find there the value of h-index, g-index and i-index of organization.

Publications of the last 5 years are analyzed in detail, including the data on publications in different journal groups (Russian journals, foreign journals, RSCI journals, etc.), their citedness in RISC, number of publications that were cited at least once, distribution of publications by OECD fields of science.
Yearly bibliometric and altmetric indicators for the last 10 years are also available together with statistics for paper distribution by science disciplines, types of documents, organizations, authors, citations, number of coauthors.

**Journal profile**

Full journal information includes fields that make possible journal search by multiple parameters: journal title, publisher, ISSN, country, subject area, language, indexed/not in Web of Science or Scopus, availability of full texts and others.

Page with analysis of journal demonstrates a set of journal metrics. The essential indicators are the number of documents and citations received during the last 8 years and various statistical reports: distribution of indexed documents and references to them by subject areas, keywords, authors, organizations, type of citing publications, et al.
Access

The access to RISC is free to all users after registration. RISC and eLIBRARY.RU are widely used not only in Russian Federation. The number of registered users and top-25 countries by registered users are demonstrated in Figures 6 and 7 respectively.

Users have free access to the full texts of open access journals and to the services like creation the sets of articles, journals and authors. Almost 50% of full-text journals in RISC are in open access. The access to the full text in the journals under subscription is available to the researchers and students from institutional network if the institution is subscribed to these journals, or for individual researchers by payment for a certain article. The registered users can make their own remarks to a document, discuss it with other users and even evaluate the publication. These remarks and evaluations are available to other users and can be seen in altmetric section in publication description.
The usage of the database is shown in Figure 8. Remarkably, the ratio of views/downloads is rather stable over time.

**Analytics services**

RISC provides several standard forms of analytical reports mentioned above — the analysis of publication output of an author, organization and journal.

The registered user may create his own publication sets. The analysis of user-created publication sets presents a summary of the total number of publications, the number of papers in academic journals, the number of articles in journals included into the Web of Science or Scopus, the number of articles in journals included in the RISC core, the number of articles in journals included in the RSCI, the average impact factor of the journals in which the articles were published, the number of authors, the average number of publications per author, the total number of citations of publications, the average number of citations per article, the number of articles cited at least once, the number of self-citations (articles from the same set), h-index of the publication set.

**Journal metrics**

Like all citation databases, RISC allows to evaluate the quality of academic journals by citation metrics. There is a whole range of different indicators calculated for this purpose.

At first, there is a set of ‘impact factor-like’ indicators that show the average number of citations per document in a certain journal. Two-year and five-year impact factors with or without journal self-citations are calculated for each journal in RISC. Additionally, the two-year impact factor is calculated which includes citations from all document types from RISC, so the share of non-journal publications in the total number of citations to journal’s papers may be assessed.

It should also be noted that the high value of the impact factor does not guarantee the high quality of an academic journal. This index can be artificially inflated by self-citations or citations from the ‘friendly’ journals. Therefore, it should be considered in conjunction with other indicators calculated in the RISC. One must pay attention to the share of journal self-citations and its Herfindahl index for citing journals. High values of these parameters (greater than 40% for the self-citation index and more than 1,500 for the Herfindahl index) indicate that a significant proportion of cites to the journal come either from its own papers or from a very limited number of other periodicals.

In eLIBRARY.RU this index is also used for organizations mentioned as an affiliation of a citing author in journal articles. It helps to assess if the scientist is cited widely, not from its own institution.
The number of journal citations and self-citations, the average number of items in the reference list, citing and cited half-life is also calculated for RISC journals.

**Science Index journal indicator**

The special complex journal indicator is calculated in RISC — Science Index journal indicator. Its methodology includes normalization by OECD subject categories, by average size of reference lists in the science field, by chronological distribution of citations and the share of them leading to papers included to eLIBRARY.RU database. It resembles the SNIP indicator (Moed, 2010), but differs by citation window (Science Index uses 5 years instead of 3 in SNIP). Additional detail is that Herfindahl-Hirschman index is added to the methodology which weights how diverse are citations the journal receives.

This methodology helps to make cross-disciplinary comparisons and rank journals more accurately.

**Journal selection for RSCI**

The aim of the joint project of the company Thomson Reuters (its scientific department which is now ‘Clarivate Analytics’) and eLIBRARY.RU is to select the best Russian academic journals and make a separate database Russian Science Citation Index (RSCI) which is available at the Web of Science platform.

![Figure 12. Journals recommended by experts and included in RSCI](image)

Selection of the best journals from RISC for RSCI database at the Web of Science platform was made by bibliometric evaluation and peer expertise. The expert council was established. The council consists of the scientists and science administrators from leading institutions (Russian
Academy of Sciences and top Universities). For the expertise top-10% of scientists with high bibliometric indicators for every discipline were selected, it means that about 30,000 experts were invited. The final choice of journals was made by a working group when both the expertise and bibliometric analysis were combined. The results are shown in Fig. 12.

Finally, 652 journals were selected for RSCI. Distribution of RSCI journals by subject compared with RISC is shown in Figure 13.

**Conclusion**

Russian journals are aggregated now at the eLIBRARY.RU platform. The first and probably the most important point is that Russian journals have now become visible. All regional journals from distant regions now are available for researchers, at minimum at the level of titles/abstracts.

One of the main functions of science is communication (Merton, 1968). As sociologists of science think, if a brilliant research is not communicated to others, this is not a scientific achievement. So the creation of eLIBRARY.RU makes Russian science much better. We may even say that from classic sociological point of view the launch of such a database often turns non-science to science in Russia. Now a researcher from Moscow has a chance to communicate with a researcher from Vladivostok (more than 6,000 km distance), this is to say be alerted about his papers and be informed of his current research published in a local Vladivostok journal. This is very important for a science in a big country. And many of these papers can be read in full text at eLIBRARY.RU.
Moreover, the platform also helps to assess citedness of journals and their scientific level. One may make reliable evaluation of the quality of indexed journals, publication activity of authors and organizations. RISC includes much more Russian journals from social sciences and humanities than can be found in international citation databases, Web of Science and Scopus.

Today RISC is widely used for the analysis of Russian institutes and universities. For example, one case is given by Zibareva and Parmon (2012). Most probably, this trend will become more pronounced in the future.

References
Arefiev P., Eremenko G. & Glukhov V. (2012) Russian Index of Science Citation – the instrument for research evaluation. Bibliosphera, 5, 66–71.
Ye J. (2014) Development, significance and background information about the "Chinese Book Citation Index" (CBkCI) demonstration database. Scientometrics, 98(1), 557–564.
Accuracy of citation data in Web of Science and Scopus

Nees Jan van Eck  Ludo Waltman

{ecknjpvan, waltmanlr}@cwts.leidenuniv.nl
Centre for Science and Technology Studies, Leiden University, Leiden (The Netherlands)

Abstract
We present a large-scale analysis of the accuracy of citation data in the Web of Science and Scopus databases. The analysis is based on citations given in publications in Elsevier journals. We reveal significant data quality problems for both databases. Missing and incorrect references are important problems in Web of Science. Duplicate publications are a serious problem in Scopus.

Conference topic
Data accuracy and disambiguation

Introduction
Citation relations between publications are the cornerstone of many bibliometric analyses. For this reason, the availability of accurate citation data is essential for high-quality bibliometric studies. In this paper, we present a large-scale analysis of the accuracy of citation data in Web of Science (WoS) and Scopus, the two most important multidisciplinary bibliometric databases. Our work continues earlier research on this topic (e.g., Buchanan, 2006; Franceschini, Maisano, & Mastrogiacomo, 2013, 2015; García-Pérez, 2010; Olensky, Schmidt, & Van Eck, 2016). The analysis that we present focuses on citations given in publications that appeared in journals published by Elsevier.

The accuracy of citation data can be studied from two perspectives. On the one hand, we can study the extent to which the reference lists of publications are properly represented in the reference data in a bibliometric database. On the other hand, we can study the degree to which a bibliometric database, based on the reference data it contains, manages to correctly identify citation relations between publications indexed in the database. The first problem is about the accuracy of reference data, while the second problem is about the accuracy of citation matching. Our primary focus in this paper is on studying the accuracy of reference data, although we will also provide some insight into the accuracy of citation matching.

Data
We used the Elsevier ScienceDirect Article Retrieval API to obtain the reference lists of all Elsevier publications that appeared in the period 1987–2016 (except for a small share of publications not included in the subscription of our university). The reference lists were obtained in XML format. We refer to the data set that we obtained in this way as the Elsevier data set. Publications without references were excluded from this data set.

For each publication in the Elsevier data set, we tried to find corresponding publications in WoS and Scopus. In the case of WoS, we focused on publications of the document types article and review published in the period 1987–2016 and indexed in the Science Citation Index Expanded, the Social Sciences Citation Index, or the Arts & Humanities Citation Index. In the case of Scopus, we considered publications of the document types article, review, and conference paper published in the period 1996–2015. Publications from 2016 were not taken into account, because most of these publications are still missing in the most recent version of the Scopus database that we have available internally at our center. Publications from before 1996 were not taken into account because of the limited coverage of these publications in Scopus.
To link publications in the Elsevier data set to publications in WoS and Scopus, we first attempted to match publications based on DOI. If no DOI-based match could be obtained, we tried to match publications based on the combination of the name of the first author, the publication year, the volume number, and the first page number. A match was required for all four fields. In the case of matching based on the name of the first author, only the last name and the first initial of the author were taken into account.

Publications in the Elsevier data set that have been linked to publications in WoS and Scopus are referred to as linked publications. The analyses presented in this paper are restricted to these linked publications. Figure 1 shows both for WoS and for Scopus the number of linked publications per year.

![Figure 1. Time trend in the number of linked publications.](image)

The next step was to match the references in the linked publications in the Elsevier data set with publications in WoS and Scopus. References in the Elsevier data set do not include a DOI. Matching was therefore done based on the combination of the name of the first author, the publication year, the volume number, and the first page number. References for which a match could be made are referred to as linked references.

Table 1 reports some statistics summarizing the results of linking the Elsevier data set with WoS and Scopus.

<table>
<thead>
<tr>
<th></th>
<th>WoS</th>
<th>Scopus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document types</td>
<td>article, review</td>
<td>article, review,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>conference paper</td>
</tr>
<tr>
<td>No. of linked publications</td>
<td>6,063,087</td>
<td>5,006,165</td>
</tr>
<tr>
<td>No. of references in Elsevier data set</td>
<td>206,540,477</td>
<td>172,125,831</td>
</tr>
<tr>
<td>No. of references in WoS/Scopus</td>
<td>203,349,407</td>
<td>170,153,108</td>
</tr>
<tr>
<td>No. of linked references</td>
<td>135,559,190</td>
<td>84,391,846</td>
</tr>
</tbody>
</table>

**Results**

*Analysis based on the number of references in a publication*

We start by making a comparison for the linked publications of the number of references in the Elsevier data set and the number of references in WoS and Scopus. In this comparison, we
do not make use of the linking of references discussed above. We consider only the number of references in a publication. We distinguish between linked publications for which the number of references in WoS or Scopus is equal to, larger than, or smaller than the number of references in the Elsevier data set. A fourth class are linked publications that have no references at all in WoS or Scopus. Table 2 reports the share of the linked publications in WoS and Scopus belonging to the different classes. Time trends are presented in Figure 2.

Table 2. Classification of linked publications based on their number of references.

<table>
<thead>
<tr>
<th></th>
<th>WoS</th>
<th>Scopus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal no. of references</td>
<td>77.2%</td>
<td>96.4%</td>
</tr>
<tr>
<td>More references</td>
<td>2.7%</td>
<td>1.2%</td>
</tr>
<tr>
<td>Fewer references</td>
<td>19.3%</td>
<td>1.2%</td>
</tr>
<tr>
<td>No references</td>
<td>0.8%</td>
<td>1.2%</td>
</tr>
</tbody>
</table>

Figure 2. Time trend in the classification of linked publications based on their number of references.

Figure 2 shows that in recent years there are no or almost no linked publications without references in WoS and Scopus. In the case of Scopus, for almost all linked publications the number of references in Scopus is the same as in the Elsevier data set. The situation is quite different in the case of WoS. A relatively small share of the linked publications have more references in WoS than in the Elsevier data set. Moreover, a large share of the linked publications have fewer references in WoS than in the Elsevier data set. To better understand why some linked publications have more references in WoS than in the Elsevier data set, we manually examined ten randomly selected cases. It turns out that a single reference in the Elsevier data set sometimes refers to multiple cited works. It seems that WoS aims to split up such a reference into multiple references, each of them referring to a single cited work. Based on the ten cases that we examined, we have the impression that the approach taken by WoS to split up references does not always give good results. In some cases, however, it works very well. We for instance found a publication (DOI: 10.1016/j.jorganchem.2015.04.038) that includes the following reference:

This is seen as a single reference in the Elsevier data set. In WoS, this reference has been correctly split up into the following three references:

GAUSS J, 1992, CHEM PHYS LETT, V191, P614
GAUSS J, 1993, J CHEM PHYS, V99, P3629
SCHAFER A, 1992, J CHEM PHYS, V97, P2571

In Scopus, the reference has not been split up. Instead, the reference has been truncated. In Scopus, it refers only to a single cited work. The other two cited works have simply been left out.

Of course, we also need to better understand why there are so many linked publications that have fewer references in WoS than in the Elsevier data set. The analysis reported below provides more insight into this issue.

Analysis based on linked references

We now present a second analysis. This analysis is based on the linked references identified using the approach described in the Data section. For each linked reference, we searched for the corresponding citation relation in WoS or Scopus. In the case of WoS, we used the citation relations identified by a citation matching algorithm developed at our center. We did not have access to the ‘official’ citation relations identified by the database producer, but our own citation relations have been shown to compare favorably with these ‘official’ citation relations (Olensky et al., 2016). In the case of Scopus, on the other hand, we did use the ‘official’ citation relations identified by the database producer.

Of the 136 million linked references in the WoS case (see Table 1), 0.9% did not have a corresponding citation relation. In the Scopus case, 1.2% of the 84 million linked references did not have a corresponding citation relation. Figure 3 shows the time trend in the percentage of linked references for which there is no corresponding citation relation in WoS or Scopus.

![Figure 3. Time trend in the percentage of linked references for which no corresponding citation relation was found in WoS or Scopus.](image)

As can be seen in Figure 3, for both databases the percentage of linked references without a corresponding citation relation has decreased significantly over time. In recent years, in the case of Scopus, about 1% of the linked references do not have a corresponding citation relation. This percentage is somewhat lower for WoS. The difference may at least partly be explained by the use of an advanced citation matching algorithm in the case of WoS.

Why do some linked references have no corresponding citation relation in WoS or Scopus? To answer this question, we manually examined a number of randomly selected linked
references for which there is no corresponding citation relation. We focused on linked references in publications that appeared in 2015. In the case of WoS, 0.7% of the linked references in publications from 2015 do not have a corresponding citation relation. Of the 125,583 linked references without a corresponding citation relation, we examined a random sample of 60 cases. It turns out that the 60 cases can be classified into the following four categories:

- **Missing reference (33 cases, 55.0%)**: The reference is missing in WoS (see also Buchanan, 2006; García-Pérez, 2010). Our examination revealed that in some cases multiple references are missing in the same publication. We found one case (DOI: 10.1016/j.virol.2015.02.016) in which the first part of the reference list of a publication is entirely missing. Interestingly, this turned out to coincide exactly with the references listed on one specific page in the PDF file of the publication, suggesting that even in 2015 indexing of Elsevier publications in WoS was still partly based on PDF files.

The problem of missing references explains why in the analysis based on linked publications discussed above we found that a significant share of the linked publications have fewer references in WoS than in the Elsevier data set.

- **Error in reference (16 cases, 26.7%)**: There is an error in the reference in WoS, such as an incorrect publication year or volume number.

- **Incorrect reference (10 cases, 16.7%)**: The reference has been replaced by a completely different reference (sometimes referred to as a ‘phantom reference’; García-Pérez, 2010). The latter reference has some similarity with the former one (e.g., name of the first author, publication year, and perhaps volume number or first page number), but apart from this it is completely different. Some examples of incorrect references are presented in Table 3. We consider incorrect references to be highly problematic, since they result in serious distortions in citation records.

- **No problem (1 case, 1.5%)**: There is no problem. A closer examination showed that for this linked reference there actually does exist a corresponding citation relation, but this citation relation had been overlooked by our algorithms.

In the case of Scopus, 1.2% of the linked references in publications from 2015 do not have a corresponding citation relation. Of the 73,598 linked references without a corresponding citation relation, we examined a random sample of 30 cases. The 30 cases can be classified into the following three categories:

### Table 3. Examples of incorrect references in WoS. Elements that the original reference and the incorrect reference in WoS have in common are shown in bold.

<table>
<thead>
<tr>
<th>Reference in WoS</th>
<th>Reference in original publication</th>
</tr>
</thead>
</table>

In the case of Scopus, 1.2% of the linked references in publications from 2015 do not have a corresponding citation relation. Of the 73,598 linked references without a corresponding citation relation, we examined a random sample of 30 cases. The 30 cases can be classified into the following three categories:
Missing reference (6 cases, 20.0%): The reference is missing in Scopus. In each of the six cases that we examined, we found that in fact the entire reference list of the citing publication is missing.

Duplicate publications (9 cases, 30.0%): The reference is available in Scopus, but there is a problem related to duplicate publications (Valderrama-Zurián, Aguilar-Moya, Melero-Fuentes, & Aleixandre-Benavent, 2015). There exist multiple records in Scopus for the publication cited in the reference. When creating citation relations, it is therefore not clear which record should be used for the cited publication. In the nine cases that we examined, we found that the Scopus citation matching algorithm had created a citation relation using an inferior record for the cited publication. The record is incomplete, while another more complete record could have been used as well.

Citation matching problem (15 cases, 50.0%): The reference is available in Scopus and a citation relation can be created, but for some reason the Scopus citation matching algorithm has failed to do so.

As mentioned above, the version of the Scopus database that we used in our analysis is not entirely up to date. We therefore also examined the 30 cases discussed above in the online version of Scopus. Interestingly, we found that 13 of the 30 cases have been corrected in the online version of Scopus, which seems to suggest that Scopus data quality has improved significantly during the past year.

Conclusions

Citation data suffers from inaccuracies both in WoS and in Scopus. However, the inaccuracies are of a quite different nature. Missing references are a quite significant problem in WoS. The problem of incorrect references is even more serious. This problem needs to be fixed urgently. In Scopus, the citation matching algorithm seems to need improvement. Moreover, duplicate publications (Valderrama-Zurián et al., 2015) represent an important data quality problem in Scopus that requires serious attention.

Our analysis has focused on citations given in publications in Elsevier journals. These publications represent a significant share of all publications indexed in WoS and Scopus. However, it is not clear to what extent our findings generalize to publications from other publishers.

References


Faculties Activity Research based on Local and Global Databases.
Case Study.

Veslava Osinska¹ Piotr Malak²

¹ wieo@umk.pl
Nicolaus Copernicus University, Torun (Poland)

² piotr.malak@uwr.edu.pl
Wroclaw University, Wroclaw (Poland)

This research is sponsored by Polish National Science Center (NCN) under grant 2013/11/B/HS2/03048: Information Visualisation methods in digital knowledge structure and dynamics study.

Introduction
The state of under-representation of the humanities and social sciences in the global databases such as Web of Science and Scopus has become permanent and detrimental. But for scientometric research on mezzo (institutional) level local databases can successfully complement world indexes. The authors present case study of using unstructured data, deriving from University Website and then visually analysing. Cleaning and processing allows to identify metadata which describe academic activities in organizing different events at University. Another data source is University bibliographic database Expertus, the most complementary information resource concerning the publications of employees. This way, two different activities: publishing and organisational one can be compared for each faculty or equivalent division and thus evaluate collective impact in academic environment.

Data and metadata

Data collecting
Application form accessible at University Website https://www.umk.pl/badania/konferencje/ allows to collect such metadata about conferences as: Date, Title, Faculty, Organiser(s), Place, Leader, Secretary, Description. Downloaded data due to HTML format required semiautomatic cleaning. The relatively small records quantity \( N = 1344 \) allowed for quick estimation of similarities and relationships. Due to field about organizers units asymmetric matrix has been created for the quantitative representation of the intensity of cooperation in conference organizing. Asymmetry is connected with unequal roles of the organizer and co-organizer.

Data aggregation
All identified University units (over 300) were grouped into the main categories corresponding to 17 Faculties and one additional, General. Thus, for publishing and organisational activity, cumulative data according 18 categories was obtained.

In order to compare the received visualisations with the generated image according to the Web of Science data, the next level of grouping: faculties into scientific areas was performed. Ultimately, into study we involved domains, such as:
- humanities,
- natural science,
- life sciences,
- medicine/clinical sciences,
- engineering and technical science,
- social sciences.

Analysis
Graphs, maps and diagrams were used in the visualisations. To illustrate the dynamics of publishing activity for each Faculty area chart is used as the most effective in the trend study (Fig.1).

![Figure 1. Dynamics of faculties publishing per year.](image)
The band width exposes publication quantity a year for each faculty. It allows to observe all changes and parallel to capture information which is the most effective unit for analyzed period.

For comparison local and global databases, ring charts were used. Figure 2 presents domain structure for three data sources. As mentioned before, standardized categorization was applied. All domains colored by scheme, which is available at InCites reports: https://incites.thomsonreuters.com/. Outer ring at Fig. 2 shows the best representation of publications for humanity and social sciences according local database –Expertus.
Figure 2. Domain structure of publications of University employees according three databases.

The visualisation of social activities working on joint ventures (conferences, symposia, seminars) can be accomplished by using graphs (Garfield, 1994). We used the legible kind of graph – circular network layout (Fig. 3). Band width indicates the intensity of cooperation in organizing conferences. This way we discovered the close relations in such kind of activity between faculties (for example history and politology).

In order to prove this observation NLP techniques were used. Texts of conferences description were processed by Ward’s clusters analysis (1963) and visualised - the result we can see on Figure 4. Dendrogram presents faculties similarity based on texts clustering. Visualisation results confirmed previous assumptions according relationships between organizational units.

The set of metadata predisposes for extending mapping and broadening the perspective of study.

We can also carry out geomapping and provide analysis of foreign cooperation of University. We can compare publishing versus mobility for each year, but all maps exceed the scope of current work.

Figure 3. Co-organizing local events by different faculties.

Figure 4. Co-organizing local events by different faculties.

Summary

The authors present analytical possibilities of local database in confrontation with global scientometric indexes. Humanity and social sciences became visible on science map based on local data. Visualisation techniques were chosen according data type and analysis purposes (Osiniska & Malak, 2016). Thus, circular graph shows cooperation between units, area chart – dynamics, ring diagram – comparison and structure and dendrogram – close and far similarity. Authors proposed publishing activity of scientists to complement by alternative one - organisational and this way to broaden evaluation framework.

References

Garfield, E. (1994) Scientography: Mapping the tracks of science. Social & Behavioural Sciences 7 (45), 5-10;
Evaluating efficiency of digital databases used for scientific production in Chinese universities

Jianpin Wang 1,2  Rongying Zhao 1,2 *

1 Research Center for Chinese Science Evaluation, Wuhan University, Wuhan, China
2 School of Information Management, Wuhan University, Wuhan, China

Abstract: It is essential for universities to make optimum use of the resources available for scientific research. The evaluation of efficiency can provide information for making evidence-based decisions. Online databases are currently the avenues by which researchers have access to the scientific-technological information required to carry out their activities. In this work, we, using Data Envelopment Analysis (DEA) and bibliometric method, evaluate the efficiency of the online databases used for scientific production in China in 2015. We choose "Project 985" universities as the sample. The DEA inputs include the number of Databases (DBs) and the number of full-time professors (FTPs), and the output concludes the number of documents indexed in Scopus. The results reveal that the use of inputs to produce scientific publications is inefficient at most of the analyzed universities. The policy recommendations are: improving the libraries’ research support service, promoting the use of open access resources, adjusting policies of incentives for research.

Keywords: Scientometrics; Bibliometrics; research evaluation; Database; Efficiency; Data Envelopment Analysis (DEA)

Conference Topic
Science policy and research assessment; Methods and techniques; Journals, databases and electronic publications

* Corresponding authors: Rongying Zhao, zhaorongying@126.com
Jianpin Wang, ggowangjianpin@126.com
Introduction

In the Information and Knowledge Society, the most important contribution by universities is the generation of Intellectual Capital (Ramezan, 2011). At the level of universities, the production of Intellectual Capital is concretized in the consolidation of networks with other stakeholders and in the production of documents (articles, books, book chapters, patents, Conferences, publishers, among other scientific productions). In order to leverage the generation of Intellectual Capital, the availability of financial capital in universities is imperative (Huggins, 2008). Since the availability of financial capital is restricted in higher education systems (HES) in developing countries and sometimes in developed countries, efficient management of economic resources is a strategic step (Abbott & Doucouliagos, 2003; Kao & Hung, 2008). Therefore, the efficiency analysis of HES has received a growing attention for the formulation of policies related to the optimized use of resources to improve quality and impact of teaching and research.

As to Chinese HES, the greatest challenge is how to build world-class universities, as well as its projection towards the dynamics of globalization driven by Information Communications and Technology (ICT). This is not only the important missions of Chinese HES in the 21st century, but also the important measure to realize the great rejuvenation of the Chinese nation. Then there are multiple constraints that constitute this problem. Two are mentioned by way of illustration. The first restriction is related to the quality of teaching and research. During the past thirty years, Chinese HES developed so fast that it has become the world’s largest higher education system. Regarding the quality of 2560 higher education institutions (technical, technological and university institutions) however, only 38 of them are “Project 985” (1.5%) and 116 are “211-project” (4.5%). At an international comparative level, according to QS University Rankings, there are only 4 Chinese universities within the list of the first 100: Tsinghua University (25), Peking University (41), Fudan University (51), Shanghai Jiaotong University (70). The second restriction is related to funding. Chinese expenditure on higher education for every student as a percentage of GDP has been declining since China began expanding university enrolment in 1999.

To summarize, In Chinese HES the weaknesses persist, new problems emerge, and the economic resources to deal with this are increasingly scarce. For these reasons, it is necessary to ascertain whether the universities are getting an optimal research output with the financial and human resources that they use. So, expanding and deepening the spectrum of efficiency studies in Chinese HES is a relevant task in different scenarios.

Studies on efficiency in HES have crossed four paths: (1) Public policy (Geva-May, 2001); (2) Universities, at the aggregate or individual level (ARCELUS & Coleman, 1997; Chen & Chen, 2011; Johnes, 2006; Klimova Kozyrev & Babkin, 2016; Shen, Wu, Business, & University, 2016); (3) Academic departments and research units (institutes, centers, groups, among others) (Agasisti, Catalano, Landoni, & Verganti, 2012; Wang & Guan, 2005); And (4) university libraries (Simon Simon & Arias, 2011; Sommersguterreichmann, 2010). As can be seen, the literature on the efficiency of HES is large, however, in the context of the Chinese HES, the studies on efficiency of digital database (DB) have not been found. Based on these studies, this work seeks to contribute to the evaluation of DB efficiency in Chinese universities in a new way.

Universities produce knowledge through research, they disseminate it by training graduates and postgraduates and by publishing the results of the research (Pastor & Serrano, 2016). The review of literature in the university efficiency studies finds two ubiquitous variables: (1) The number of teachers or researchers; And (2) Research results or documents.

On the other hand, we identify the absence of an extremely important variable for the current context of scientific research: ICT (Wang, Carley, Zeng, & Mao, 2007; Wuchty Jones & Uzzi, 2007). According to the World Bank, ICT is the hardware set, Software, networks and media.
used for the collection, storage, processing, transmission, and presentation of data, information, knowledge, services, and applications. ICT is restructuring the institutional fabric of higher education and influencing the academic work done by university teachers. Higher education institutions are not only producing and supporting technological innovations but are at the same time intensive users and subject to the limitations of ICT (Välimaa & Hoffman, 2008). Therefore, this study will focus on DB, which are currently the avenues by which researchers have access to the scientific-technological information required to carry out their activities.

With the above, this study has the objective of evaluating the efficiency in the use of the DBs consulted for the scientific production by the "Project 985" universities. Building on the advances identified in the literature, this work contributes on three fronts: (1) An overview of the inventory of the DBs of "Project 985" universities is presented; In the search of literature, a similar contribution was not found, therefore, this study advances significantly in this sense; (2) It presents a relative evaluation of efficiency in the use of these DBs by full-time teachers using DEA method. On one hand, to our knowledge, there is no article devoted to the efficiency evaluation of universities’ DB in China. On the other, the DEA technique has long been utilized as an invaluable tool for conducting efficiency evaluation, but its combination with bibliometric method has been relatively slow; And (3) An updated description of the scientific production of documents indexed in Scopus by "Project 985" universities during the period of 1996-2015 is presented.

In short, this study has the potential to help management determine proper policies and be used as a tool to optimize the investment of Chinese universities for research, since the relative decision-making entities do not currently have the tools to make such evaluation.

The study is organized as follows. After this introduction, Section 2 describes the methodology, the sample of universities studied, and the additional sources consulted. Section 3 very briefly presents the results analysis. The last section (Section 4) discusses the study's conclusions, makes some recommendations, and offers suggestions for future research.

Methodology

Data Envelopment Analysis

The traditional economist definition of efficiency is: The optimal use of inputs to achieve the greatest output (Cook & Seiford, 2009). In this study, the efficiency is measured empirically using Data Envelopment Analysis (DEA), a mathematical optimization model introduced by Charnes et al. (CharnesCooper & Rhodes, 1978) that generalized Farrell’s single input/output measure into a multiple input/multiple-output technique. DEA model estimates the convex hull of the production function by a linear programming optimization, and the inefficiency is defined by a convex envelope of the data.

In our analysis, the extension CCR will be applied. This extension has two orientations: (1) Products (CCR-Outputs); Or (2) Inputs (CCR-Inputs). The first orientation seeks to maximize results with the same inputs. The second orientation seeks to reduce inputs to maintain results. This study follows the first orientation.

Charnes et al. (1978) proposed that the efficiency measure of every DMU is obtained by obtaining the maximum of a ratio between results and weighted inputs subject to the condition that the similar proportions of each DMU should be less than or equal to unity. In this way, the definition of the classical ratio of engineering is generalized: single-output / single-input to a form of multiple-outputs / multiple-inputs, without requiring a priori weight assignment. In this sense, the model is formulated formally:

$$\max h_0 = \frac{\sum_{k=1}^5 u_k y_{k0}}{\sum_{i=1}^m v_i x_{i0}}$$

1097
\[
\text{S.T. } \frac{\sum_{k=1}^{s} u_k y_{kj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1; j = 1, \ldots, n. \\
v_k, v_i \geq 0; k = 1, \ldots, s; i = 1, \ldots, m.
\]

Where \( y_{kj}, x_{ij} \) (all positive) are the outputs and inputs of the DMUs analyzed \((j)\); And that \( u_k, v_i \geq 0 \) are the weighted variables to be determined when solving the problem. The efficiency of each of the members of the DMU \((j = 1, \ldots, n)\) is evaluated relatively to the others. Thus, the efficiency index extracted for each DMU is 0 to 1, and the DMU with index 1 are those that delimit the border of efficiency.

**Sample and data**

The DMU sample defined for this study are the 38 Chinese "Project 985" universities (Minzu University of China is not concluded because no documents generated by this universities indexed in Scopus)( shown in Table 1). These universities were selected for four reasons. The first, because these universities generate about 44.4% of the national scientific production. The second, because most teachers with doctorates in the country are linked to these institutions. The third, because these universities are most likely to become the world-class university. The fourth, because most of national investment in Research and Development (R&D) is concentrated in these universities.

The inputs defined for the DEA are two: (1) the number of DBs (digital databases of the library system of each university); And (2) the number of full-time professors (FTPs) (See Table 1). To know the number of DBs and FTPs, we go to the web site of each university of the sample. This investigation took place in the middle of 2016. The number of DBs and FTPs all correspond to 2015. These two inputs were defined because at present the DBs are the most consulted sources by the FTPs, being these the ones in charge of carrying out activities of investigation, teaching, and extension.

The output defined for the DEA are the documents generated by the "Project 985" universities and indexed in Scopus in 2015 (See Table 1). Scopus was chosen because it is the most complete indexed documents database in the market. They took into account the documents of all disciplines (Health Sciences, Life Sciences, Physical Sciences and Social Sciences). All types of documents (articles, reviews, articles in the press, books, book chapters, conference articles, conference reports, letters, editorials, notes, short articles, business articles or articles and errata) were also taken into account.

<table>
<thead>
<tr>
<th>Chinese &quot;Project 985&quot; universities</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DB</td>
<td>FTP</td>
</tr>
<tr>
<td>Tsinghua University</td>
<td>176</td>
<td>1355</td>
</tr>
<tr>
<td>Shanghai Jiaotong University</td>
<td>191</td>
<td>858</td>
</tr>
<tr>
<td>Zhejiang University</td>
<td>205</td>
<td>1552</td>
</tr>
<tr>
<td>Peking University</td>
<td>247</td>
<td>2221</td>
</tr>
<tr>
<td>Harbin Institute of Technology</td>
<td>102</td>
<td>975</td>
</tr>
<tr>
<td>Huazhong University of Science and Technology</td>
<td>180</td>
<td>1115</td>
</tr>
<tr>
<td>Sichuan University</td>
<td>110</td>
<td>1611</td>
</tr>
<tr>
<td>Tongji University</td>
<td>190</td>
<td>930</td>
</tr>
<tr>
<td>Jilin University</td>
<td>164</td>
<td>2016</td>
</tr>
<tr>
<td>Fudan University</td>
<td>183</td>
<td>1016</td>
</tr>
<tr>
<td>Shandong University</td>
<td>113</td>
<td>1082</td>
</tr>
<tr>
<td>Xi'an Jiaotong University</td>
<td>105</td>
<td>836</td>
</tr>
<tr>
<td>Sun Yat-Sen University</td>
<td>188</td>
<td>1390</td>
</tr>
<tr>
<td>Beihang University</td>
<td>74</td>
<td>583</td>
</tr>
</tbody>
</table>
Results

In the first instance, a regional picture is presented on the performance of China in terms of scientific production indexed in Scopus during the period 1996-2015 (Table 2). Subsequently, a description of the national affiliation of the authors of the documents for 2015 (Figure 1) is presented. Then, an overview of the documents is presented according to its discipline and category for 2015 (Figure 2 and Figure 3). Finally, the results of the DEA are presented (Table 3 and Figure 4). In this vein, Table 2 presents the top ten countries in the scientific production ladder in the world.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Documents</th>
<th>Citable documents</th>
<th>Citations</th>
<th>Self-citations</th>
<th>Citations per document</th>
<th>H index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>United States</td>
<td>9360233</td>
<td>8456050</td>
<td>202750565</td>
<td>94596521</td>
<td>21.66</td>
<td>1783</td>
</tr>
<tr>
<td>2</td>
<td>China</td>
<td>4076414</td>
<td>4017123</td>
<td>24175067</td>
<td>13297607</td>
<td>5.93</td>
<td>563</td>
</tr>
<tr>
<td>3</td>
<td>United Kingdom</td>
<td>2624530</td>
<td>2272675</td>
<td>50790508</td>
<td>11763338</td>
<td>19.35</td>
<td>1099</td>
</tr>
<tr>
<td>4</td>
<td>Germany</td>
<td>2365108</td>
<td>2207765</td>
<td>40951616</td>
<td>10294248</td>
<td>17.31</td>
<td>961</td>
</tr>
<tr>
<td>5</td>
<td>Japan</td>
<td>2212636</td>
<td>2133326</td>
<td>30436114</td>
<td>8352578</td>
<td>13.76</td>
<td>797</td>
</tr>
<tr>
<td>6</td>
<td>France</td>
<td>1684479</td>
<td>1582197</td>
<td>28329815</td>
<td>6194966</td>
<td>16.82</td>
<td>878</td>
</tr>
<tr>
<td>7</td>
<td>Canada</td>
<td>1339471</td>
<td>1227622</td>
<td>25677205</td>
<td>4699514</td>
<td>19.17</td>
<td>862</td>
</tr>
<tr>
<td>8</td>
<td>Italy</td>
<td>1318466</td>
<td>1217804</td>
<td>20893655</td>
<td>4825002</td>
<td>15.85</td>
<td>766</td>
</tr>
<tr>
<td>9</td>
<td>India</td>
<td>1140717</td>
<td>1072927</td>
<td>8458373</td>
<td>2906102</td>
<td>7.41</td>
<td>426</td>
</tr>
<tr>
<td>10</td>
<td>Spain</td>
<td>1045796</td>
<td>966710</td>
<td>14811902</td>
<td>3510196</td>
<td>14.16</td>
<td>648</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Documents</th>
<th>Citable documents</th>
<th>Citations</th>
<th>Self-citations</th>
<th>Citations per document</th>
<th>H index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td></td>
<td>1045796</td>
<td>966710</td>
<td>14811902</td>
<td>3510196</td>
<td>14.16</td>
<td>648</td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td>2716785</td>
<td>2515419.9</td>
<td>44727482</td>
<td>16044007.2</td>
<td>15.14</td>
<td>878.3</td>
</tr>
<tr>
<td>SD</td>
<td></td>
<td>2212636</td>
<td>1857761.5</td>
<td>27003510</td>
<td>7273772</td>
<td>16.34</td>
<td>829.5</td>
</tr>
<tr>
<td>Max</td>
<td></td>
<td>2378328.96</td>
<td>2273963.6</td>
<td>56822526.1</td>
<td>27828543.05</td>
<td>5.08</td>
<td>372.76</td>
</tr>
<tr>
<td>Min</td>
<td></td>
<td>9360233</td>
<td>8456050</td>
<td>202750565</td>
<td>94596521</td>
<td>21.7</td>
<td>1783</td>
</tr>
</tbody>
</table>

As shown in Table 2, the total number of the Chinese scientific documents indexed in SCOPUS was 4,076,414 during the year of 1996-2015, ranking in the second place while the United States was 9,36,233, more than two-fold as China, ranking in the first place. On the other hand, during the year of 1996-2015, the number of the Chinese citations was 24,175,067, ranking in the seventh place in the world, and the number of Citations per document was 5.93 times, far below the world average level which was 14.16; Another aspect to highlight is the Chinese H-index (563), which is only higher than India in these ten countries. All this suggests that China has become a large country of scientific production, while there is still wide gap in quality compared with the advanced countries.

Source: Scopus, 2015.

**Figure 1. National affiliation of the authors of the documents in 2015**

Another aspect to highlight is the international collaboration and exchange in science and technology have been expanded. As shown in Figure 1, in 2015 there are 444,804 documents indexed in Scopus where at least one of the authors has affiliation to a Chinese entity, followed by the United States with 41,109 and the United Kingdom with 9,314. From this it can be inferred that Chinese partners in scientific production are United States and United Kingdom. There is also substantial collaboration with Australia (3th place) and Hong Kong (4th place).

Source: Scopus, 2015.

**Figure 2. Production of documents by discipline**

As shown in Figure 2, the discipline with the greatest number of documents was engineering with 143,667 (32.3%), followed by medicine with 91,645 (20.6%), and material science with 1,046 (14.8%). On the other hand, the production in social sciences, being in the categories of
others, was not displayed separately. Thus, it is possible to infer that more than 65.9% of Chinese scientific production in 2015 was generated in natural science disciplines such as engineering, medicine, material science, physics and astronomy, chemistry, computer science.

Source: Scopus, 2015.

Figure 3. Production by type of documents

As can be seen in Figure 3, the largest number of documents corresponds to the category of articles with 365,593 (82.2%), followed by conference articles with 58,802 (13.2%), and reviews with 9,681 (2.2%).

Table 3 presents the efficiency index. It is important to emphasize that the efficiency index extracted from the DEA for each DMU is from 0 to 1, with DMUs with index 1 that delimit the efficiency frontier. As shown in Table 3, the University with the most DB is Peking University with 247. The University with the least DB is Renmin University of China with 70. The average is 140 DB, then 39.47% of the sample universities are above this average. The University with the most FTP is Peking University with 2,221. The University with the least FTP is National University of Defense Technology with 320. The average is 945, then 36.84% of the universities in the sample are above this average. The University with the largest production of documents is Tsinghua University with 12,630. The University with the lowest production is Renmin University of China with 894. The average number of documents is 5,197, and 50% of the sample are above this average. In summary, the universities of the sample present an asymmetry between the availability of inputs (DB and FTP) and outputs (documents). This is evident from the wide standard deviations.

On the other hand, the mean of the efficiency index (0.575) is higher than the median (0.545). This means that the distribution of the data is not symmetric and positively biased. Thus, universities with a higher efficiency rate than the average are less frequent in the sample. The two universities that bordered the efficiency frontier were Shanghai Jiaotong University and Beihang University (5.3% of the sample). Two universities obtained an efficiency index higher than 0.90, Harbin Institute of Technology and Tsinghua University (5.3% of the sample). 11 universities (28.9% of the sample) obtained an efficiency index in the range of 0.6 to 0.8. Most universities, equivalent to twenty-three (60.5% of the sample), are less than or equal to 0.6. (See Figure 4).

In summary, a polarization in the efficient use of DB by FTP for the production of documents is evidenced. While 2 (5.3%) of the universities in the country make efficient use of resources for the advance of scientific production, the efficiencies in most of the universities are quite low. In other words, the relationship between human resources and scientific-technological
information resources, and the production of documents, is weak. In addition, there is no efficient use of these resources in general.

<table>
<thead>
<tr>
<th>Chinese &quot;Project 985&quot; universities</th>
<th>Input</th>
<th>Output</th>
<th>efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DB</td>
<td>FTP</td>
<td>Documents</td>
</tr>
<tr>
<td>Shanghai Jiaotong University</td>
<td>191</td>
<td>858</td>
<td>10801</td>
</tr>
<tr>
<td>Beihang University</td>
<td>74</td>
<td>583</td>
<td>5805</td>
</tr>
<tr>
<td>Harbin Institute of Technology</td>
<td>102</td>
<td>975</td>
<td>7687</td>
</tr>
<tr>
<td>Tsinghua University</td>
<td>176</td>
<td>1355</td>
<td>12630</td>
</tr>
<tr>
<td>Sichuan University</td>
<td>110</td>
<td>1611</td>
<td>6589</td>
</tr>
<tr>
<td>University of Science and Technology of China</td>
<td>138</td>
<td>547</td>
<td>5231</td>
</tr>
<tr>
<td>National University of Defense Technology</td>
<td>98</td>
<td>320</td>
<td>3041</td>
</tr>
<tr>
<td>Xi'an Jiaotong University</td>
<td>105</td>
<td>836</td>
<td>6149</td>
</tr>
<tr>
<td>Tianjin University</td>
<td>102</td>
<td>762</td>
<td>5651</td>
</tr>
<tr>
<td>Beijing Institute of Technology</td>
<td>127</td>
<td>515</td>
<td>4633</td>
</tr>
<tr>
<td>University of Electronic Science and Technology</td>
<td>163</td>
<td>445</td>
<td>3956</td>
</tr>
<tr>
<td>Shandong University</td>
<td>113</td>
<td>1082</td>
<td>6150</td>
</tr>
<tr>
<td>Southeast University</td>
<td>116</td>
<td>794</td>
<td>5732</td>
</tr>
<tr>
<td>Zhejiang University</td>
<td>205</td>
<td>1552</td>
<td>10709</td>
</tr>
<tr>
<td>South China University of Technology</td>
<td>93</td>
<td>1738</td>
<td>4766</td>
</tr>
<tr>
<td>Dalian University of Technology</td>
<td>111</td>
<td>741</td>
<td>4677</td>
</tr>
<tr>
<td>Tongji University</td>
<td>190</td>
<td>930</td>
<td>6485</td>
</tr>
<tr>
<td>Huazhong University of Science and Technology</td>
<td>180</td>
<td>1115</td>
<td>6973</td>
</tr>
<tr>
<td>Northwestern Polytechnical University</td>
<td>131</td>
<td>544</td>
<td>3747</td>
</tr>
<tr>
<td>Hunan University</td>
<td>128</td>
<td>475</td>
<td>3243</td>
</tr>
<tr>
<td>Chongqing University</td>
<td>125</td>
<td>743</td>
<td>4440</td>
</tr>
<tr>
<td>Fudan University</td>
<td>183</td>
<td>1016</td>
<td>6178</td>
</tr>
<tr>
<td>China Agricultural University</td>
<td>70</td>
<td>598</td>
<td>2859</td>
</tr>
<tr>
<td>Nanjing University</td>
<td>161</td>
<td>992</td>
<td>5573</td>
</tr>
<tr>
<td>Central South University China</td>
<td>131</td>
<td>943</td>
<td>4927</td>
</tr>
<tr>
<td>Jilin University</td>
<td>164</td>
<td>2016</td>
<td>6472</td>
</tr>
<tr>
<td>National University of Defense Technology</td>
<td>120</td>
<td>528</td>
<td>3235</td>
</tr>
<tr>
<td>Peking University</td>
<td>247</td>
<td>2221</td>
<td>9178</td>
</tr>
<tr>
<td>Sun Yat-Sen University</td>
<td>188</td>
<td>1390</td>
<td>6029</td>
</tr>
<tr>
<td>Wuhan University</td>
<td>202</td>
<td>1260</td>
<td>5747</td>
</tr>
<tr>
<td>Lanzhou University</td>
<td>75</td>
<td>595</td>
<td>2439</td>
</tr>
<tr>
<td>Nankai University</td>
<td>109</td>
<td>711</td>
<td>2792</td>
</tr>
<tr>
<td>Xiamen University</td>
<td>149</td>
<td>787</td>
<td>3304</td>
</tr>
<tr>
<td>Beijing Normal University</td>
<td>154</td>
<td>862</td>
<td>2961</td>
</tr>
<tr>
<td>Northwest A&amp;F University</td>
<td>133</td>
<td>548</td>
<td>1937</td>
</tr>
<tr>
<td>Ocean University of China</td>
<td>120</td>
<td>567</td>
<td>1793</td>
</tr>
<tr>
<td>East China Normal University</td>
<td>131</td>
<td>1737</td>
<td>2076</td>
</tr>
<tr>
<td>Renmin University of China</td>
<td>194</td>
<td>612</td>
<td>894</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Input</th>
<th>Output</th>
<th>efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>139.711</td>
<td>944.842</td>
<td>5197.079</td>
</tr>
<tr>
<td>Median</td>
<td>131</td>
<td>815</td>
<td>5079</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>41.464</td>
<td>463.079</td>
<td>2586.490</td>
</tr>
<tr>
<td>Max</td>
<td>247</td>
<td>2221</td>
<td>12630</td>
</tr>
<tr>
<td>Min</td>
<td>70</td>
<td>320</td>
<td>894</td>
</tr>
</tbody>
</table>

1102
Conclusions

The generation of knowledge on the part of the universities is a pillar strategy for the progress of the Information and knowledge Society. In the case of China several restrictions are identified that impede the progress of the higher education system in this regard, including restrictions on quality, and budget. In this sense, it is essential to make optimum use of the resources available for scientific research. However, the availability of inputs that provide information for making evidence-based decisions is restricted. This study seeks to contribute to this gap through an evaluation of efficiency in the use of DB for scientific production by the "Project 985" universities, using DEA and bibliometric method.

China has performed relatively well for it is the second largest country in the quantity of scientific production in the world, while there is still wide gap in quality compared with the advanced countries.

With regard to the co-authors of documents generated by researchers, the majority is in the United States and United Kingdom, above the collaboration with Australia.

As for the areas of knowledge, one of the greatest production is engineering, followed by medicine, and material sciences. Faced with document types, most focus on scientific articles, followed by conference articles, and reviews.

Regarding to the evaluation of efficiency in the use of database for scientific production, its horizon is asymmetric. In addition, the average of FTP of the 38 universities is 945, of which, only 36.84% are above this level. This exhibits an obvious concentration. Faced with ownership in the number of BD, only 39.47% of the sampled universities are above average (140). With regard to document production, only 50% of the sampled universities are above average (5197). Regarding the efficiency index, only 44.7% of the universities in the sample were above average (0.575).

As a point of reference, two of the 38 universities in the sample efficiently use both databases and their research staff for document generation. Finally, it was observed that more than 60% of the 38 universities are below 0.6 of efficiency, where universities with sufficient resources have the performance that do not agreed the efficiency frontier. In other words, the relationship between human resources and scientific-technological information resources, and the production of documents, is weak. Moreover, there is no efficient use of these resources in general.
From this perspective, this paper suggests three recommendations. The policy recommendations are: improving the libraries’ research support service, promoting the use of open access resources, and adjusting policies of incentives for research.

The first recommendation aims to deploy alternative strategies to help the researchers in university achieve eResearch, which refers to the use of information technology to support existing and new forms of research. Some Australian universities have already made some headway there. In order to create a good environment, the information management policy and development trend need to be understood and analyzed; respects of personnel, the Libraries should explore their tasks, the change of their role and the researchers’ demand, so as to become partners of academic community; at the technical level, the libraries should, based on the system platform, concentrate on appropriately describing and effectively uncovering the data, which can promote the research collaboration and innovation (Wilkinson & Treloar, 2008).

The second recommendation suggests that it is necessary to consider the trend and the impact of open access scientific literature. Open access scientific literature allows free, online and peer-reviewed scientific production online. According to the Joint Information Systems Committee (JISC, 2008), open access literature is an alternative for institutions that can not acquire high-priced databases. For its part, the United Nations Educational, Scientific and Cultural Organization (UNESCO) at the World Open Educational Resources Congress recommended that all Member States implement freely available resources at both levels of formal and informal education (Hoosen, 2012).

The third line of recommendations expresses the need to formulate alternative policies for research incentives. According to Franzoni et al. (Franzoni & Stephan, 2011), publication incentive policies involving money bonuses increase the number of articles submitted for review but not the acceptance rate for publication. In fact, the acceptance rate has a negative correlation with the incentive policies based on money bonuses. Additionally, it is proposed that the incentives associated with the academic career possibilities are those that actually have a correlation with the quantity and quality of published research. These policies implemented in countries such as Spain or Germany make access to higher educational levels, promotion, and salary dependent on the performance of researchers, rather than the publication bonus.

In summary, the perspectives on the growth of the scientific production in China are not conclusive. While there was a notable performance in the past which has made the country a visible place in the world, it is necessary to strengthen research capacities in universities as a whole, not only the “Project 985” universities. Faced with the current scenario on the use of DB by FTPs for document production, most universities have to reevaluate their relative performance and propose alternative strategies, such as the improvement of libraries’ research support service, the use of free resources, and adjustment of research incentive policies. Future research in this area could increase the sample of higher education institutions as a whole, not just the "Project 985" universities. They could also take into account the impact of the documents, not only on research documents but also on patents, in order to trace the impact of research on the productive sector.

Acknowledgments
This paper is supported by National Social Science Foundation in China (Grant No.16BTQ055). The authors would like to thank Prof. Junping Qiu who is in Research Center for Chinese Science Evaluation for his valuable comments and suggestions to improve the quality of the manuscript.

References:

About the Author

Jianpin Wang is a PhD Candidate, majoring in information science. Her research interests focus on knowledge management and competitive intelligence, information measurement and scientific evaluation. Contact: School of Information Management, Wuhan University, Wuhan, 430072, China, Email: ggwangjianpin@126.com.
Rongying Zhao is a Professor at the School of Information Management, Wuhan University. Her research interests focus on knowledge management and competitive intelligence, information measurement and evaluation. Rongying Zhao is the corresponding author.
and can be contacted at: School of Information Management, Wuhan University, Wuhan, 430072, China, Email: zhaorongying@126.com.
Abstract
Research funding based on scientific metrics is strongly required; however, navigating the funding landscape is challenging. Conventional techniques that use intercitation and co-citation cannot be utilized for ongoing projects. Therefore, to grasp research projects on grants and visualize their relationship, this paper proposes to locate projects in a multi-dimensional space using a paragraph vector technique with an entropy-based clustering method. After describing our novel method, we provide details of the experiments we conducted to show that the proposed method successfully formed meaningful project clusters from 25,607 project descriptions from the 7th Framework Programme. Finally, we identify significant relationships from the project map, e.g., projects with similarities in different subject categories and projects bridging different categories.

Conference Topic
Knowledge discovery and data mining, Mapping and visualization

Introduction
Information about research projects is important to investigators, policy analysts, advocacy organizations, and funding agencies. Thus, there is a need to globally analyze research projects funded by grants and visualize their relationships. However, current quantitative approaches to understanding research activities focus on what authors tell us about past accomplishments through analysis of published research papers, and thus most maps of science from literature characterize what was accomplished after the fact. Therefore, this paper focuses on what researchers want to work on and creating a map characterizing what is currently being attempted on research projects in order to provide new insights.

Most funding agencies have their own categories or taxonomy to classify projects. However, even if two projects are assigned the same category, relationships, such as any distance and semantics, may not be found, and projects may be multifaceted and thus have several overlapping categories. Funding agencies also use different categories, and there is no comprehensive scheme to characterize projects; thus, projects cannot be compared between different agencies.

Automatic topic extraction from research projects using lexical approaches, such as Probabilistic Latent Semantic Analysis (pLSA) (Steyvers & Griffiths, 2007) and Latent Dirichlet Allocation (LDA) (Blei, Ng & Jordan, 2003), has been investigated. However, the relationships between projects cannot be measured directly. Moreover, research project descriptions are limited and do not include attributes, such as citations and references. Thus, techniques using intercitation and co-citation cannot be utilized, although projects will eventually include articles in their research results.

Recently, natural language processing (NLP) techniques, e.g., word/paragraph embedding, have been proposed to find relationships between unstructured documents. Such embedding techniques represent words and paragraphs as real-valued vectors of several hundred dimensions. Then, the types and degrees of the semantic relations among the vectors are calculated by the similarities between vectors. This paper proposes a method to extract relationships between research projects using word/paragraph vectors. In addition, to address
an unclustered problem associated with paragraph vectors, we introduce an entropy-based clustering method for word vectors to create concept vectors.

The remainder of this paper is organized as follows. Section 2 discusses related work, and after the introduction of the dataset Section 3 describes a baseline method and the proposed method for extracting relationships between research projects. Experiments and evaluations are described in Section 4, and conclusions and suggestions for future work are given in Section 5.

**Related Work**

As described previously, funding agencies and publishers generally have their own classification systems. In the 7th Framework Programme for European Research and Technological Development (FP7), most projects have three-digit subject index classification (SIC) codes that represent academic subjects, and some projects have more than one code. Thus, interdisciplinary projects can be found by searching multi-labelled projects; however, it is difficult to find similar projects in a simple manner. Moreover, comparison of articles with Association for Computing Machinery (ACM) classification (https://www.acm.org/publications/class-2012) or Springer Nature classification requires taxonomy exchanges.

Previous studies have examined automatic topic classification based on content, such as documents and articles. One uses LDA (Griffiths & Steyvers, 2004) to find the five most probable words in the topics, and each document is viewed as a mixture of topics. Thus, this approach can classify documents across different agencies and publishers. However, the relationship between any pair of projects/articles, such as similarity and correlation, cannot be computed directly.

In this regard, the NIH Visual Browser (Talley et al., 2011; Herr et al., 2009) (http://nihmaps.org/index.php) computed the similarity between projects as the mixture of classification probability to each topic based on pLSA, using the average symmetric Kullback-Leibler divergence function (Kullback & Leibler, 1951). However, this similarity is a combination of probabilities, i.e., it is not derived from actual content semantics.

In another study, a hierarchically structured set of topics with Springer Nature Classification tags was created from Springer Nature Proceeding papers based on an ontology (Osborne et al., 2016). However, the sources were limited to computer science and the similarity or distance between pairwise articles could not be computed.

In a broader sense, mapping projects funded under the FP7 has been visualized based on articles by project members in the Web of Science bibliographic database by examining co-authorship and co-citation patterns (Salah et al., 2013). In addition, Maps of Science (http://mapofscience.com/) is a website that provides the Sci2Tool visualization tools (Borner, 2013) and maps of journals and documents (Boyack, Klavans & Borner, 2005). Here the similarity between journals is calculated using some metrics, such as cosine and Jaccard similarity, based on the attributes of journals (including intercitation and co-citation). Note that these maps promote interdisciplinary research collaboration but do not directly represent project relations.

In contrast, a word/paragraph vector, which is a distributed representation of words and paragraphs, is attracting attention in NLP. Assuming that context determines the meaning of a word (Firth, 1957), words appearing in similar contexts are considered to have similar meaning. In the basic form, a word vector is represented as a matrix whose elements are the co-occurrence frequencies between a word $w$ with a certain usage frequency in the corpus and words within a fixed window size $c$ from $w$. A popular representation of word vectors is word2vec proposed by Google (Mikolov et al., 2013a; Mikolov et al., 2013b). Word2vec creates word vectors using
a two-layered neural network obtained by a skip-gram model with negative sampling. Specifically, word vector is obtained by calculating the maximum likelihood of objective function $L$ in Eq. 1, where $T$ is the number of words with a certain usage frequency in the corpus. Word2vec clusters words with similar concepts in a vector space.

$$L = \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j}|w_t)$$ (1)

Additionally, Le and Mikolov (2014) proposed a paragraph vector that learns fixed-length feature representations using a two-layered neural network from variable-length pieces of texts, such as sentences, paragraphs, and documents. A paragraph vector is considered another word in a paragraph and is shared across all contexts generated from the same paragraph but not across paragraphs. The contexts are fixed-length and sampled from a sliding window over the paragraph. The paragraph vectors are computed by fixing the word vectors and training the new paragraph vector until convergence, as shown in Eq.2.

$$L = \sum_{t=1}^{T} \log p(w_t|w_{t-c}, ..., w_{t+c}, d_i)$$ (2)

, where $d_i$ is a vector for a paragraph $i$ that includes $w_t$. While word vectors are shared across paragraphs, paragraph vectors are unique among paragraphs and represent the topics of the paragraphs. Thus, by considering word order, paragraph vectors can also address the weakness of bag-of-words models in LDA and pLSA. Therefore, paragraph vectors are considered more accurate representations of the semantics of the content. As a result, we can input vectors directly into analysis using machine learning and clustering techniques to find similar projects/articles in different academic subjects and the relationships between projects/articles from different agencies and publishers.

**Extraction of Project Relationships**

This section introduces our experimental dataset and then describes a baseline method and the proposed method to extract relationships between research projects.

**Funding description dataset**

In this paper, we analyzed project information from FP7 obtained from the Community Research and Development Information Service (CORDIS) website (http://cordis.europa.eu/fp7/) for the following reasons.

- FP7 includes several research areas, from bioscience to computer science.
- Most projects have been completed; thus, the entire structure is included.
- Projects are classified by more than 70 subject terms, i.e., SIC codes.

Precisely, our experimental dataset consists of the titles and descriptions of 25,607 FP7 projects from 2007 to 2013, including 305,819 sentences in total. All words in the sentences were tokenized and lemmatized before creating the following vector space.

**Paragraph Vectors as Baseline**

Before introducing the proposed method, this section presents a problem in applying the paragraph vectors for research project descriptions.
We implemented the paragraph embedding technique using the Deep Learning Library for Java (https://deeplearning4j.org). Then, we constructed paragraph vectors for 25,607 FP7 projects. The hyperparameters were set empirically through experiments as follows; 500 dimensions for 66,830 words that appeared more than five times, the window size \( c \) is 10, and the learning rate and the minimum learning rate are 0.025 and 0.0001 with an adaptive gradient algorithm (AdaGrad), respectively. The learning model is a Distributed Memory model (PV-DM) with hierarchical softmax.

In Fig. 1, each dot represents a project, and the distance between projects is the cosine similarity of the paragraph vectors. As a result, we found that projects are scattered and not clustered by any subject or discipline in the vector space. Most projects are connected to a small number of projects. Thus, it is difficult to grasp trends and compare an ordinary classification system, such as SIC codes. Looking closely at the vector space, some reasons for this unclustered problem are that each word with nearly the same meaning has slightly different word vectors, and shared but unimportant words are also considered the commonality of paragraphs. In fact, Le and Mikolov (2014) stated that classification accuracy with multiple categories was less than 50%. Thus, to address this problem, we introduce an entropy-based clustering method for the word vectors before constructing the paragraph vectors.

**Paragraph Vectors with Entropy-based Clustering of Word Vectors**

This section proposes a method for constructing concept clusters of word vectors, aiming to make paragraph vectors meaningful groups from the scientific and technological point of view. The fact that synonyms tend to gather in a word vector space indicates that the semantics of a word spatially spread to a certain distance. This observation is also suggested in related literature (Vilnis & McCallum, 2015). Thus, to unify word vectors of the same concept, excluding trivial common words, we generated concept vectors of the word vectors based on the significance of each concept. The proposed flow is as follows. We first extracted 19,685...
concepts (broader terms) with more than one hyponyms (narrower terms) from the JST science and technology thesaurus. The JST thesaurus primarily consists of keywords that have been frequently indexed in 36 million papers accumulated by the JST since 1975. Currently, this thesaurus includes approximately 276,179 terms with English and Japanese notations in 14 categories from bioscience and computer science to civil engineering. The JST thesaurus is publicly accessible from Web APIs in the J-GLOBAL website (http://jglobal.jst.go.jp/en/), along with the visualization tool Thesaurus Map (http://thesaurus-map.jst.go.jp/jisho/fullIF/index.html (in Japanese)). We then calculated the information entropy (Shannon, 1948) of each concept in the FP7 projects. Next, after creating concept clusters according to the degree of entropy, we unified all word vectors in the same concept cluster to a concept vector and constructed paragraph vectors based on the concept vectors. The overall flow is shown in Fig. 2.

Figure 2. Construction of paragraph vectors based on concept vectors.

Shannon's entropy in information theory is an estimate of event informativeness. Thus, we used this entropy to measure the semantic diversity of a concept that is an idea borrowed from Santus et al., (2014).

Figure 3. Concepts in the JST thesaurus.
As shown in Fig. 3 and Eq. 3, we calculated the entropy of a concept \( C \) considering the appearance frequencies of a hypernym \( T_0 \) corresponding to the concept and its hyponyms \( T_1 \ldots T_n \) as an event probability. The frequencies of synonyms \( S_{i0} \ldots S_{im} \) of term \( T_i \) were summarized to a corresponding concept. Here, \( p(S_{ij}|C) \) is the probability of word \( S_{ij} \) given concept \( C \) and term \( T_i \). For each concept in the JST thesaurus, we calculated the entropy \( H(C) \) in the FP7 dataset. As the probabilities of events become equal, \( H(C) \) increases. If only particular events occur, \( H(C) \) is reduced due to low informativeness. Thus, the proposed entropy of a concept increases when a hypernym and hyponyms that construct a concept appear with a certain frequency in the dataset, respectively. Therefore, the degree of entropy indicates the semantic diversity of a concept. Then, assuming that the degree of entropy and the spatial size of a concept in a word vector space are proportional to a certain extent, we split the word vector space into concept clusters. In fact, our preliminary experiment indicated that the entropy of a concept has high correlation \( R = 0.602 \) with the maximum Euclidean distance of hyponyms in the concept in a vector space, at least while the entropy is rather high. Specifically, we modified a clustering method called \( x \)-means (Pelleg & Moore, 2000) to refine clusters by repeatedly subdividing them until the defined criterion was satisfied. In our method, we set the determination condition as shown in Eq. 4. This condition represents that the vector space \( w_0 \ldots w_n \) is subdivided according to the ratio of two words \( w_i, w_j \) with the highest entropies \( E(i), E(j) \) in a cluster. Each cluster is subdivided until entropy becomes under 0.25 (threshold of the top 1.5%) or the number of elements in a cluster is less than 10. These parameters were also determined empirically through experiments.

\[
C(w) = \begin{cases} 
C(w_i) & (E(i) > \frac{||w_m - w_i||}{||w_m - w_j||}) \\
C(w_j) & \text{(otherwise)}
\end{cases}
\]

(4)

, where \( C(w) \) is a concept vector of \( w \). After generating 1,260 concept clusters from 66,830 word vectors, we considered the centroid of all vectors in a cluster as a concept vector. Then, we constructed paragraph vectors based on the concept vectors rather than from each word vector, as shown in Eq. 5.

\[
L = \sum_{t=1}^{T} \log p(C(w_t)|C(w_{t-c}), ..., C(w_{t+c}), d_t)
\]

(5)

The above process is expected to have the following effects. Using concepts with high entropy included in the JST thesaurus, which are significant in scientific and technological contexts, as common points with each paragraph vector (excluding trivial words), paragraph vectors can comprise meaningful groups in the vector space. Simultaneously, unknown synonyms and closely related words not defined in the thesaurus can be included in the concept vector if they are in the same cluster. Taking the centroid vector as a representative vector in a cluster is to separate each concept vector as much as possible to form a clear difference in the vector space.

**Experiments and Evaluation**

The Distributed Recursive Graph Layout (DrL) algorithm, which produces edge-weighted force directed graphs, was used to visualize the relationships between projects. The inputs to the DrL
algorithm are pairwise similarities, where the weight of each edge in the network represents the similarity between two adjunct projects. The DrL algorithm produces a layout wherein similar nodes are placed close to one another and shows the local clustering of related nodes and the global structure of related clusters. As shown in Fig. 4, we computed 328 million cosine similarities for all pairs of the 25,607 projects; however, we only kept those that were above a given threshold as edges.

Figures 1 and 4 show results obtained with the same hyperparameters for the baseline and proposed methods, respectively. We confirm that the proposed method (Fig. 4) has several clusters in the paragraph vectors compared to the baseline method (Fig. 1). For more quantitative comparison, Figure 5 shows the relationships between the cosine similarities and the number of edges, and the relationship between degree centrality and the number of nodes (i.e., projects) in the case of the cosine similarities of >0.35. As a result, we confirmed that
edges with higher cosine similarity and nodes with higher degrees increase. Therefore, note that the numbers of projects appeared in Figs. 1 and 4 are different when the figures have the same edge threshold.

Figure 5. Comparison between paragraph vectors and ones with entropy clustering.

For qualitative evaluation, we compared the map with the categories of the FP7 projects, i.e., SIC codes. Then, we focused on the middles of clusters, which are centers of research projects, the edges of clusters, which are novel projects because of the longest distance from the center, and also bridges in a gradient between clusters. These bridge projects provide information about the relationships and interdisciplinary approaches between different subjects and also suggest potential areas for future projects for funding agencies, since new clusters may be created on the bridges in future.
Here, we describe four example findings from the obtained funding map. In Fig. 6 corresponding to area A1 in Fig. 4, the middle cluster is the biggest cluster including more than 5,000 projects, which correspond approximately 20% in the map. Although we need an in-depth analysis for the inside of the cluster, we confirmed that there are three obvious bridges to the clusters for Biotechnology (BIO) and Medical Biotechnology (MED), Social Aspect (SOC), and Information and Media (INF) and Information Processing and Information Systems (IPS).

In Fig. 7 corresponding to area A2 in Fig. 4, in terms of fusion of humans and machines, projects about robots and human sensors in the INF and related studies in the BIO and the MED, e.g., inserting chips into the human brain, are located in the left and right clusters. Then, sociological and ethical research in the SOC runs across the two clusters as a bridge. This may suggest the possibility of future collaboration of such studies. In addition, although we need a consensus of the novelty, there seem to be novel projects on the bridge, such as ‘seeing’ with the ears and several biomimetic studies. We will further investigate the extraction of such novel projects. Moreover, by comparing 2008 and 2014 industrial projects, such as Industrial Biotech (IBI) and Industrial Manufacture (IND) such projects increased six-fold in the second half of the FP7, and we confirmed that the subjects were initially limited but finally spanned a wide range across the map.

In contrast, some projects with the Life Science (LIF) code that were obviously about Aerospace Technology (AER), were in fact clustered as AER projects in the map. This case supports the certainty of the map.

Finally, we are evaluating the accuracy of the map with a statistical sampling method. In our preliminary experiment, we randomly extracted 100 pairs of the projects with the cosine similarity of >0.5 so as to make the distribution similar to the one of all pairs. Each pair has two project titles and descriptions, and a cosine value that is divided into three levels: weak (0.5 <= cos. < 0.67), middle (0.67 <= cos. < 0.84) and strong (0.84 <= cos.). Then, three members of our organization, a funding agency in Japan confirmed the similarity of each pair. The members were provided the prior explanations for the intended use of the graph and some examples of evaluation. As a result, we confirmed that 78% of the project similarities (i.e., the distances in the graph) matched majority votes of the members’ opinions. By contrast, the accuracy of the
distances in Fig. 1 was 21%. The evaluation results were determined to be in “fair” agreement (Fleiss’ Kappa = 0.29).

Conclusion and Future Work

Navigating the funding landscape is challenging. However, the research projects do not have references; thus, analyzing the projects using journal citation linkages has difficulty. Therefore, we assessed project relationships using recent semantic similarity measures. After improving the existing paragraph vector technique with an entropy-based clustering method, we successfully formed lattice-like clusters, which are linked by strings of aligned projects whose subjects are relevant to each other. Finally, we confirmed the good face validity and retrieved some relations of particular interest from the map.

As the next step, we need to extract new insights from the proposed map of research projects. Especially, we will compare our map to previous science maps based on published research papers, since they could be applied to assess whether projects are novel by extracting conventional topic words from clusters in the citation-based maps and/or journal disciplines and calculating the similarity between the topic words and project descriptions. In future, we could use the paragraph vectors for several statistical applications, such as multi-dimensional scaling and clustering, rather than drawing cosine similarity networks. In particular, we could consider extracting more semantic relations of projects, such as the most common topics between two projects, from the offset vectors of the projects. We plan to perform additional comparisons with other project datasets, such as FP7 and JST funding projects. Since concept vectors can be shared across different languages, English documents and Japanese documents can be compared through the JST thesaurus with English and Japanese notations. Furthermore, visualizing overlaid heterogeneous datasets, such as funding projects and conference papers, will be of interest, as well as the visualization of their temporal changes to detect emerging research areas. The current FP7 funding maps is publicly accessible in our website (http://togodb.jst.go.jp/sample-apps/map_FP7/).

References


LitStoryTeller: An Interactive System for Visual Exploration of Scientific Papers Leveraging Named entities and Comparative Sentences

Qing Ping\textsuperscript{1} Chaomei Chen\textsuperscript{2}
\textsuperscript{1}qp27@drexel.edu
Drexel University, Philadelphia (America)
\textsuperscript{2}cc345@drexel.edu
Drexel University, Philadelphia (America)

Abstract
The present study proposes \textit{LitStoryTeller}, an interactive system for visually exploring the semantic structure of a scientific article. We demonstrate how \textit{LitStoryTeller} could be used to answer some of the most fundamental research questions, such as how a new method was built on top of existing methods, based on what theoretical proof and experimental evidences. More importantly, \textit{LitStoryTeller} can assist users to understand the full and interesting story a scientific paper, with a concise outline and important details. The proposed system borrows a metaphor from screen play, and visualizes the storyline of a scientific paper by arranging its characters (scientific concepts or terminologies) and scenes (paragraphs/sentences) into a progressive and interactive storyline. Such storylines help to preserve the semantic structure and logical thinking process of a scientific paper. Semantic structures, such as scientific concepts and comparative sentences, are extracted using existing named entity recognition APIs and supervised classifiers, from a scientific paper automatically. Two supplementary views, ranked entity frequency view and entity co-occurrence network view, are provided to help users identify the “main plot” of such scientific storylines. When collective documents are ready, \textit{LitStoryTeller} also provides a temporal entity evolution view and entity community view for collection digestion.

Conference Topic
Mapping and visualization; Knowledge discovery and data mining; Methods and techniques.

Introduction
With the sheer volume of scientific publications every year, it becomes a double-challenge for researchers to not only comprehend a collection of research articles as a whole, but also to grasp effectively important pieces of information scattered everywhere in each single article. As a solution to this double-challenge, researchers from multiple areas have contributed insights. In the domain of scientific mapping, some existing work have proposed applications to digest a collection of research papers on collection-level, such as CiteSpace (Chen, 2006), Action Science Explorer (Dunne, Shneiderman, Gove, Klavans, & Dorr, 2012), VOSViewer (Van Eck & Waltman, 2010). In broader scope of digital humanity, several applications have been developed to digest a text corpus on topic-level, such as VarifocalReader (Koch, John, Wörner, Müller, & Ertl, 2014), Serendip (Alexander, Kohlmann, Valenza, Witmore, & Gleicher, 2014), on sentence-level, such as PICTOR (Schneider et al., 2010), and on word-level, such as POSvis (Vuilleumot, Clement, Plaisant, & Kumar, 2009) and Wordle (Viegas, Wattenberg, & Feinberg, 2009).

Existing work mentioned above are insufficient to solve the double-challenge for scientific paper digestion. First, scientific mapping applications focus on extracting collection-level patterns as a whole, and are not suitable for individual document analysis. Second, applications in digital humanity, though on multiple-levels, are not tailored for scientific paper digestion. Most existing work in this area are designed for special text corpus, such as poem, play, news, Bible, and so on, but very few if not none are tailored for scientific papers. Third,
even for those applications that are not confined to one type of text, the toolkit developed for
detailed investigation is still simplified. To bridge this gap, we present LitStoryTeller for better support of scientific paper digestion. On document-level, LitStoryTeller automatically extracts scientific concepts (or entities exchangeably in this paper) from full-text and visualizes entities and their co-occurrence and comparative relations in storylines. Here we use the visual metaphor of “storyline” in a play, where entities are considered as “characters”, and paragraph/sentence are seen as “scenes” where “characters” get on stage. With this visual metaphor, we are able to preserve the logical plot of a scientific paper. Moreover, this storyline is synchronized with a text viewer, so that user could navigate through the full-text using the “characters” and “scenes” in the storyline as anchors. Supplementary views are also provided to help users to get focused on the main plot of the storylines. On collection-level, LitStoryTeller visualizes all entities in a collection with two different views, i.e. entity community view and temporal entity evolution view. To our best knowledge, this paper is among the first work that is designed to support document-level exploration using a storyline visual metaphor and leveraging a variety of techniques such as entity extraction and comparative sentence classification methods. The main contributions of the present work are as follows:
1. We develop a framework for document-level exploration of scientific papers, using a “storyline” visual metaphor that preserves the logical thinking plot of a scientific paper;
2. We develop modules for named entity recognition and comparative sentence classification that could run in real-time to support semantic-level exploration of a scientific paper;
3. We also support collection-level exploration of scientific papers, using techniques for community detection algorithm and temporal network visualization.

The rest of the paper is organized as follows. Related work discusses existing work; System design presents the design of the proposed system; Case study demonstrates the utility of the system through a case study; Conclusion draws conclusion of the present study.

Related Work

Single-document visualization
In digital humanity, previous work has been done to facilitate users’ exploration of a single document. Depending on the granularity of the visualization, applications can be divided into the following categories.

Topic-level document visualization
To facilitate exploration of a document, some applications focus on finding the latent topics of a document first, and use topics as an intermediary between words and full document for visualization. Varifocal-Reader (Koch et al., 2014) uses text segmentation method to segment full text into topical segments, and annotates entities such as person and location. Serendip (Alexander et al., 2014) uses statistical topic models as a bridge between words, topics and documents, and visualize the three elements in matrices. The advantage of this approach is that it helps to capture the topical structure of a document for easier digestion. However, it might also suffer from loss of finer-level details, such as detailed information in sentences and entities.

Sentence-level document visualization
Other applications focus on organizing and visualizing a document on sentence-level. One application chooses not to display all sentences plainly, but rather to display sentences using a fish-eye view so that salient passages will be highlighted as focal, and the rest will be put as receded (Correll, Witmore, & Gleicher, 2011). Another application extracts quote sentences from news narratives and could support searching of quotes by speakers (Schneider et al., 2010). The strength of this approach is that finer-level details (sentences) could be organized
and shown to users. On the other hand, the weakness is that sentences alone cannot bear semantic meanings, unlike a topic or an entity.

**Word-level document visualization**

There are multiple applications on word-level document visualization. One application finds frequent word usage patterns and highlights them in full-texts (Don et al., 2007). Another application supports to visualize all neighbouring words of a given word query in a word cloud view (Vuillemot et al., 2009). One work visualizes the word frequency distributions over the narrative scope (Clement, Plaisant, & Vuillemot, 2009). Another work proposes to visualize words in a document as word cloud, known as “Wordle” (Viegas et al., 2009). One work, specially tailored for play script, visualize characters-scenes as a matrix, with character-on-stage-scene as highlights (Wilhelm, Burghardt, & Wolff, 2013). There are also some works on phonetic-levels, often tailored for poem analysis (Abdul-Rahman et al., 2013; McCurdy, Lein, Coles, & Meyer, 2016).

The advantage of visualization on word-level is that it reserves the finest-level of details. However, most of these works do not reserve the relationships between words, or entities. Besides individual weaknesses, work mentioned above are not tailored scientific paper exploration. First, most of the work above are confined in specific corpus such as poem, play, news, Bible, and so on. Second, few of the existing work on word-level focus on word-word relationships. Third, existing work lacks support for semantic information extraction, such as comparative sentence classification.

**Scientific fields evolution visualization**

Our present work is also analogous to the research of visualization of scientific field evolution on an abstract level. Research in this direction usually constructs a network of concepts or keyphrases by various proximity metrics based on co-word analysis, and then clusters the concepts into scientific fields. Then temporal patterns are investigated, such as the emergence and recombination of each scientific field in the network over time (Chavalarias & Cointet, 2013), and interactions between academic push and technological pull for theories (Callon, Courtial, & Laville, 1991). Here each scientific field is considered a “unit” in the storyline of scientific evolution, whereas in our present study, the basic “unit” of the storyline is a concept within a single scientific paper. Also, a link in scientific evolution represents the high-level connection between two scientific fields, whereas in our study, a link represents the sentence/paragraph-level co-occurrence of two concepts.

**Argumentation visualization**

Our work may also be overlapped with research in argumentation visualization. Research in this area attempts to visualize the structures of argumentations, usually in an interactive collaborative learning environment, to support decision making (Kirschner, Buckingham-Shum, & Carr, 2012). Our work instead attempts to visualize the structures of concepts in a scientific paper via automatic natural language processing of the full text and interactive visualizations.

**Storyline visualization**

There are two classical works in storyline visualization. One work proposes to visualize the storyline of a screen play by arranging characters and scenes over time (Tanahashi & Ma, 2012). More specifically, each character is represented by a curved line, and each scene is emphasized by bundling all character lines of this scene closer. Another work improves on the previous one by optimizing several objective functions to make the storyline more compact and visually-pleasing (Liu, Wu, Wei, Liu, & Liu, 2013). The present study utilizes a similar design of storyline as the bare-bone template (Elvery, 2017), which use curved line to represent a character, and a rectangle with lines passing through it to represent scenes. Character lines are arranged by first grouping characters using community detection algorithm, and then place each group as far apart as appropriate.
**Named entity recognition**

Majority of existing work on named entity recognition are supervised methods. Models used include Hidden Markov Models (HMM) (Bikel, Miller, Schwartz, & Weischedel, 1997), Maximum Entropy Models (ME) (Borthwick, 1999) and Conditional Random Fields (CRF) (McCallum & Li, 2003). Supervised named entity recognition usually have superior performances. However, these methods usually require a large amount of human-labeled data in a specific domain. In the present paper, we take advantage of the Microsoft Entity Linking API (https://www.microsoft.com/cognitive-services/en-us/entity-linking-intelligence-service), which not only recognizes named entities under a wide range of topics based on Wikipedia coverage, but also links entities of variant forms together.

**Comparative sentence classification**

Bing Liu has worked on the topic of comparative sentence extraction in a series of papers. In one paper, he proposes to use manual keyword list and frequent sequence mining, together with supervised classifier to classify sentences (Jindal & Liu, 2006a). In another work, he further proposes to not only classify sentences into comparatives/non-comparatives, but also extract the subjects as well as comparative relations from comparative sentences (Jindal & Liu, 2006b). In the present paper, we implemented the full pipeline for comparative sentence classification as in (Jindal & Liu, 2006a).

**System design**

In this section, we first provide an overview of the workflow of the proposed system, and then describe each module in detail.

**Overview of system workflow**

The overview of system workflow is depicted as in Figure 1. The workflow starts with a paper-uploading page. The uploaded full-text is tokenized, POS tagged, stemmed and converted to a feature vector. Then sentence feature vectors are sent to the comparative-sentence-classifier (1 in Figure 1) to generate comparative/non-comparative labels. In the meanwhile, POS tagged and stemmed sentences are sent to the named-entity-recognition module (2 in Figure 1). The recognized entities are sent to the single-document-storytelling module (3 in Figure 1). This module utilizes the labels generated in (1) and entity-sentence-paragraph alignments generated in (2) to create a storyline visualization, together with two supplementary views. Finally, when multiple papers have been uploaded into the system, the collective-documents exploration module (4 in Figure 1) can retrieve all entity co-occurrences stored in the local repository and visualize the entire document collection with two views. We will discuss each module below in detail.

**Comparative-sentence-classification module**

A comparative sentence expresses an ordering relation between two sets of entities with respect to some common features (Jindal & Liu, 2006b). Existing packages such as nltk.corpus.reader.comparative_sents (Pantone, 2017) cannot be used in the present study because it can only work on specific corpus with labeled comparative/non-comparative classes, entity names and relations. Instead, in the current system, we built a comparative-sentence-classification pipeline by implementing and modifying algorithms proposed in paper (Jindal & Liu, 2006a).
Full-text pre-processing
The objective of pre-processing is to convert the raw full-texts into POS tagged and stemmed sentences with corresponding indexing in paragraphs. After removing irregular characters and segmenting full-text into paragraphs by line breaks, we segment each paragraph to sentences with the NLTK PunktSentenceTokenizer. In the meanwhile, we record sentence offsets, namely starting and ending positions in paragraphs.

Feature extraction. The feature extraction step aims to pre-mine all frequent sequence patterns emerging from the training corpus. A frequent sequence pattern is a sequence pattern with sufficient confidence and support in the corpus, where $t_i$ is POS tag of a surrounding word to a keyword. For example, assume the sentence "X outperforms Y" is prevalent in a corpus, then the corresponding frequent sequence pattern will be $<\{NN\}, \{VBZ, outperform\}, \{NN\}>$, where outperform is the keyword. More specifically, we perform the following steps to extract frequent sequence patterns:

(1) Construction of keyword-list. In the original work by (Jindal & Liu, 2006a), to identify comparative sentences, three categories of keywords are proposed, namely adjectival/adverbial comparatives, single-verb keywords, and phrase-keywords, 83 keywords in total. In our study, we added four keywords: "fail", "gain", "over" and "contrast".

(2) Extraction of candidate sequences. For each keyword matched in a sentence, we extract a sequence with certain window size for it as a candidate sequence. For example, for the following sentence, where outperform is in our keyword-list:

"The concatenated features A outperform the original feature set of B."

The corresponding candidate sequence will be (window size = 3):

$[(\{JJ\}), (\{VBZ\}), (\{DT\}), (\{outperform\}, \{NN\}), (\{DT\}), (\{JJ\}), (\{NN\})]$  

(3) Frequent sequence pattern mining. We adopt the PrefixSpan algorithm to mine frequent sequence patterns from all candidate sequences generated in the last step. The PrefixSpan algorithm utilizes projection of search space into prefix sequences to reduce the number of candidate subsequence generations (Han et al., 2001). In our study, we implemented the PrefixSpan algorithm, with the minimum support $\tau$ for frequent sequence set to be 0.1, and the minimum confidence set to be 0.6.
**Classifier training.** After getting all the frequent sequence patterns, we consider them as features of a sentence, and train a Bayes Classifier based on given labels (Jindal & Liu, 2006a). If a sentence satisfies one frequent sequence pattern, the corresponding feature will have value 1, otherwise 0. We manually labelled 286 sentences from research papers, and feed the training corpus into a classifier. Here we report the results of Bayes classifier, with higher precision than those of SVM and Logistic Regression classifier. The accuracy of 5-fold cross-validation is $(0.84 \pm 0.02)$ for the Bayes classifier.

**New sentence prediction.** In the prediction-phase, each sentence is first POS tagged and stemmed. Then the sentence is converted to a feature vector indicating what frequent sequence patterns this sentence satisfies. Last, the feature vector is fed into the trained Bayes classifier to generate a prediction, namely comparative or non-comparative.

**Named-entity-recognition module**
We refer entities as important scientific concepts or terminologies discussed in a research paper. To recognize such entities in a researcher paper on-the-fly with high recall is a challenging task. In the current system, we take advantage of the Microsoft Entity-Linking API. This API not only recognizes a wide range of scientific concepts recorded in Wikipedia, but also automatically performs entity disambiguation.

One down-side of using the Microsoft Entity Linking API is that there is a limit for the service per day. Therefore, we propose a batch mechanism to minimize the number of calls as few as possible for each paper. We batch sentences into blocks before sending them using API. The block size is set to be 10000 characters. Then we use the returned results from the API to map the entity offsets in the block back into its offsets in sentences and paragraphs.

**Single-Document Story Telling**
This part of the system focuses on supporting users to grasp the important pieces of detail information in a research paper through storyline visualization. First, to help users started in the storyline, we provide two ways to determine which entities to read first (Figure 1(a) and (b)). Following these leads, user could read the storylines of these entities through the “Text-storyline cross-reference” view (Figure 1(c)) iteratively and spot important scientific conclusions with the help of “comparative sentence indicators”.

**Ranked entity frequency view**
This view provides an intuitive way is to determine which entities to explore first: to look at how many times an entity has been mentioned in a document as in (Figure 1(a)). Users could start their exploration focusing on these top-ranked entities.

**Local entity co-occurrence network view**
We could also consider the importance and interestingness of an entity to be its authority over other entities as in view (Figure 1(b)). We use the force-layout(Bostock, 2017) to draw the co-occurrence network of entities. The size of the entity indicates its frequency in a document; while the width of link indicates how frequently two entities co-occur in one sentence.

**Text-storyline cross-reference” view**
After determining which entities are of high priority to explore first, users could use the “text-storyline cross-reference” view for “close-reading”. More specifically, we design two views within the “text-storyline cross-reference” view, namely, the “Storyline view” and “Text view” (Figure 1(c)). The collaboration between the two views will help users to navigate through the full text following accumulative clues in the storyline viewer.
We borrow the “storyline” metaphor from theatre play for our visualization of a research paper. As in traditional “storyline” of a novel or a play, there are characters and scenes, and there are beginning, development, turning point, climax and conclusion in the plot.

A research paper shares similarity with the formality of a play. Essentially, a research idea is demonstrated through developments of concepts throughout its abstract, introduction, related work, methodology, experiment, discussion, conclusion and references. These sections can be regarded as grand scenes at a coarse-level. Further, each paragraph and each sentence can be considered as major scenes at fine-level. In our study, the scientific concepts are referred as entities, which can be extracted automatically from each sentences, paragraphs, and sections.

As in Figure 2(a), the storyline of a research paper comprises of entities (characters) and sections/paragraphs/sentences (scenes). The storyline should be read from left to right, as its development in the research paper. More specifically, the glyphs in the storyline represents different elements as follows:

1. **Entities.** Each entity, in a unique colour, is a scientific concept, represented by a curved line called lifeline, from left to right, indicating the period that the entity is first/last mentioned. The width of the curve line is in proportion to the frequency of an entity.

2. **Grande scenes.** Sections, such as abstract, introduction, or conclusion, are considered as grand scenes. These grand scenes are visualized with vertical lines separating the storyline into sections.

3. **Major scenes.** Each major scene is a paragraph or sentence with at least one recognized entity in it. It is represented by a transparent rectangle with soft corners. Each entity lifeline

---

**Figure 2. Storyline of a research paper**

(a) Initial storyline of a research paper

(b) Re-drawn storyline with focus-on-demand Cartesian fisheye view after mouse move
may pass through one or more major scenes, indicating occurrences in these major scenes. Multiple entity lifelines passing through the same major scene indicate their co-occurrences.

4. **Comparative sentence indicators/each scene.** At the bottom of the storyline, we visualize every scene with a rectangle at equal steps. The shade of the rectangle represents our confidence whether this scene contains “comparative statements”.

As in Figure 2(b), the storyline visualization is the main vehicle for users to perform various operations and navigate through the full-text document. We have enabled the following interactive operations on the storyline viewer:

1. **Focus-on-demand fisheye view.** To avoid too long and sparse a storyline, a focus-on-demand zoom-in/zoom-out mechanism is desirable. We support this goal by implementing a fish-eye effect on the storyline (Figure 2(b)-(1)). This enables users to expand any part of a storyline to see details at focus, and shrink the rest of the storyline as background.

2. **Select one or more entities.** The label as the starting point of each entity is selectable (Figure 2(b)-(2)). When clicked, the corresponding life-line of entity will be highlighted.

3. **Select a major scene.** Each major scene, represented with rectangle, is selectable (Figure 2(b)-(3)). When clicking the major scene, the “Text view” will synchronically jump into and highlight the corresponding section of the scene (Figure 1-(1)).

4. **Select a comparative statement.** Each scene is displayed at the bottom with equal steps (Figure 2(b)-(4)), whose shade indicates how likely the scene contains a comparative statement. These glyphs are also selectable and synchronized with the text viewer.

5. **Switch scene granularity.** The granularity of the scene can be switched from “paragraph” to “sentence” by clicking on the “visualize sentences” button (Figure 2(b)-(5)).

**Text view**

We enable the text view to have synchronized responsive behaviour in accord with operations in the storyline view. That is, when corresponding operations be performed in the storyline viewer, such as selecting an entity, selecting a scene (paragraph/sentence), the corresponding content within the text view will be highlighted and brought to the focus area.

**Overview of a collection**

We provide two types of visualizations: temporal entity evolution view and entity community view.

**Temporal entity evolution view**

The temporal entity evolution view is designed to visualize how entities in an entire document collection evolve over time. Some entities appear very early in the collection, marked by the publication date of its corresponding paper, while others emerge at much later stage, with strong connections to the earlier ones. These connections can also be seen as mapping between research fronts (novel entities) and intellectual base (old entities) (Chen, 2006). We take co-occurrences between old and novel entities on sentence-level as the connections, and visualize these connections chronologically to show entity evolution.

As in Figure 1-(4), we use an arc-diagram layout to visualize entity evolution over time. In our arc diagram, each circle represents an entity with label below it. Its colour indicates the corresponding paper where this entity first appeared in. Each arc connecting two circles represents co-occurrences of two entities in one or more sentences, whose thickness indicates the frequency of the co-occurrences. The entities are arranged from older to new horizontally based on their corresponding papers’ publication time where they first appeared in.
**Entity community view**

The entity community view is designed to visualize the communities within the co-occurrence network of entities across the collection. In entity communities, some entities appear as outliers, while others connect multiple other entities to form a cohesive community, or occupied important “positions” in the network such as connecting two major communities as pivotal points (Chen, 2004). These entities are worth further analysis.

As in Figure 1-③, we implement a force-layout graph (Bostock, 2017) to visualize the entire entity co-occurrence network. The force-layout attempts to minimize the number of crossings of edges in a network by optimizing energy functions(Kobourov, 2012). In our visualization, each circle represents an entity, and its size indicates its frequency in entire document collection. Each link represents co-occurrences between two entities, whose width indicates how often the two entities co-occurred together. We use the Louvain method for community detection which optimizes global modularity through updating local communities(Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). Entities belonging to one community are coloured with a unique colour.

**Case study**

We report a case study to demonstrate the purpose and functions of the current system design.

**Use Scenario**

We present the use case of a typical user named Alice. Alice is a graduate student, who has previous experience in research, and has some exposure to interactive visual analytic systems. Alice is surveying on research papers related to the topic of “Latent Dirichlet Allocation” (LDA), a probabilistic graphical model (Blei, Ng, & Jordan, 2003). Alice has already obtained a collection of full-text research papers on this topic (13 papers). Alice realizes that there are multiple models before and after the proposal of LDA, such as LSI, pLSI, hLDA and HDP. Alice is especially interested in how each model was built on top of one or more previous models, by improving the weakness of the previous ones on various metrics. Alice is also curious about what accompanying algorithms are frequently used together with these topic models, such as inference methods. These research questions are summarized as follows:

1. How is each novel model built or different from old models? Taking the example of LSI and pLSI, how does the new model (pLSI) is superior to the old model (LSI)?
2. What mathematical formula and inference algorithms are usually used together with each topic model?
3. Overall, what concepts (topic models, mathematical formula, and inference algorithms) are discussed the most in this document collection? How do these concepts evolve over time?

**Single-Document Storytelling**

To answer research question (1) and (2), Alice selects the paper of probabilistic Latent Semantic Indexing (pLSI) to see how pLSI is built on top of Latent Semantic Indexing (LSI). Alice uses the single-document storytelling view for this task.

Due to space limitation, we skip the step of finding important entities with our supplementary views, and assume Alice already identifies the most important entities. Recall that question (1) is “How is each novel model built or different from old models?” To answer this question, Alice decides to read the storylines of the two main characters: pLSI and LSI at paragraph-level. Alice observes that the two highlighted lifelines have several crossings (co-occurrences in a paragraph), as in Figure 3. Alice believes that these crossings are crucial to answer question (1). Interestingly, all crossings yield important information about the relationships between pLSA and LSI, as can be read from the tooltip of the crossings in the figure. These
tooltips clearly stated how pLSA is built on top of LSI theoretically and mathematically.

Figure 3. Storyline of "Probabilistic Latent Semantic Indexing" at paragraph-level.

Now Alice wants to go deeper into details, so she switches to read the storyline at sentence-level, as in Figure 4. This time, Alice notices multiple crossings at the “6. Experimental Results” section. Alice decides to read the sentences in these crossings. Interestingly, all crossings yield “conclusive” information about the relationships between pLSI and LSI, as can be read from the tooltips. Also, most of the crossings belong to the scenes with darker shade at bottom, indicating that pLSI and LSI are frequently compared in the section of “Experimental Results”. At this point, Alice is very confident about knowing how pLSA is different and most likely an improvement from LSI, with both theoretical foundations and experimental results.

Figure 4. Storyline of "Probabilistic Latent Semantic Indexing" at sentence-level.

To answer research question (2), Alice examines crossings in Figure 4 again. Alice finds out that LSI is mostly associated with “Singular Value Decomposition”, and pLSI with “Expectation-Maximization Algorithm”. Indeed, to solve LSI, the most common used method is “Singular Value Decomposition”, while for pLSI, the algorithm for inference is “Expectation-Maximization Algorithm”.

Collective-document exploration view

To answer research question (3), Alice uses the entity community view and temporal entity evolution view to understand relations among all the entities discussed in this collection.
Identifying communities
Under the Entity Community View as in Figure 5(a), Alice identifies major communities of sub-topics around LDA topic model in this document collection: LDA models (green), LSI-related models (blue), HDP models (light-brown). Some small communities, such as those in yellow and red are discussed only a few times, thus are more marginalized in the network.

Tracing entity evolutions
Under the Entity Evolution View, as in Figure 5(b), Alice can see how entities evolve over time. Alice pays special interests to entities with thicker arcs across papers. These entities have been frequently discussed both in its original papers and in newer papers, and this has led to a sequence of topic model names. From this sequence, Alice has a general idea of what are some older models (LSI, pLSI), what are newer models (LDA, HDP), and which new model is evolved around which old model. This provides a concise answer to question (3). At this point, Alice successfully answers the three research questions by using the current system.

Conclusion
The present study proposes LitStoryTeller, an interactive system for visually exploring the semantic structures of scientific papers. With a screen play metaphor, the proposed system provides interactive storyline of a research article, taking advantages of a variety of techniques such as named entity recognition and comparative sentence classification. Visualizations at collection-level are also provided to aid overall reading digestion. A comprehensive case study demonstrated the usefulness of the proposed system, by answering realistic research questions from research papers. There are two limitations in the current study. First, the Microsoft API may not be able to identify very recent scientific concepts not yet included in Wikipedia. Although we have provided the function of entity customization for users, we plan to develop our own scientific entity recognizer in the future. Second, the metrics of entity importance used in present study are relatively simple, namely entity frequencies and graph centrality. In the future, we plan to propose more robust metrics such as tf*idf, to eliminate entities that are very common yet not rich in meanings in their contexts.

Acknowledgments
This study is supported by the NSF project “A Visual Analytic Observatory of Scientific Knowledge” (NSF 1645264).
References


University rankings in computer science: a study and visualization of ‘geo-based’ impact and conference proceeding (CORE) scores

Philip Ehret¹, Alesia Zuccala², and Bela Gipp³

¹philip.ehret@uni-konstanz.de, ³bela.gipp@uni-konstanz.de
University of Konstanz, Konstanz (Germany)
²a.zuccala@hum.ku.dk
University of Copenhagen, Copenhagen (Denmark)

Abstract
This is a research-in-progress paper concerning two types of institutional rankings, the Leiden and QS World ranking, and their relationship to a list of universities’ ‘geo-based’ impact scores, and Computing Research and Education Conference (CORE) participation scores in the field of computer science. A ‘geo-based’ impact measure examines the geographical distribution of incoming citations to a particular university’s journal articles for a specific period of time. It takes into account both the number of citations and the geographical variability in these citations. The CORE participation score is calculated on the basis of the number of weighted proceedings papers that a university has contributed to either an A*, A, B, or C conference as ranked by the Computing Research and Education Association of Australasia. In addition to calculating the correlations between the distinct university rankings and the separate ‘geo-based’ versus CORE scores, we are in the process of developing a geographical visualization tool that presents the metrics so that they may be examined in an explorative way.

Conference Topic
University policy and institutional rankings; Science communication; Mapping and visualization

Introduction
University rankings have rapidly become an influential tool in government and educational policymaking (The Guardian, 2013), and after the first Academic Ranking of World Universities (ARWU) was introduced in 2003 (also known as the ‘Shanghai Ranking’), alternative rankings began to appear. These include, though in no specific order, the QS World Ranking of Universities, the Times Higher Education (THE) Ranking, SCImago ranking, and the Leiden Ranking. In past years, each ranking has been touted, examined for their disadvantages and advantages, and assessed on the bases of their similarities (e.g., Aguillo et al., 2010; Waltman et al., 2012; Bornmann et al., 2013). Certain methodological approaches have also been more heavily criticized than others (Liu & Cheng, 2005; Liu et al., 2005; VanRaan, 2005). Some critics, for example, are sceptical about approaches that rely on the amalgamation and use of weighted variables (Billaut, 2010), and others are concerned with reproducibility (Docampo, 2012). Most researchers tend to agree; however, that normalization is key to producing rankings, especially when using publication and citation data with variable field differences (López-Illescas et al., 2011).

University rankings are here to stay, and since can often be influential, there are always ample reasons to examine them further. One approach is to focus on statistical shortcomings within the ranking itself, but another is to identify and examine new variables of interest and test them to see if they show some form of positive, negative, or neutral relationship to that ranking. In this sense, correlation measures are useful, if they provide further information and
insight into the communication practices of a research field, where a ranking might not present the full picture. The field that we have chosen to investigate is information and computer science, and the type of university rankings that we use in this study implement either a field-normalized approach – i.e., the Leiden Ranking – or a field-specific approach, such as the QS World University Ranking.

Currently, our research is ‘in-progress’ and it is based on two components. The first component is metric in nature, and the second involves developing a visualization tool that will allow users to explore our results geographically.

With data pertaining to: a) the Leiden University Ranking (2016), b) the QS World University Ranking in Computer Science (2016), c) university-to-university directed citation counts collected from Web of Science (WoS) journal articles (2012-2016), the first part of our study will focus on the following:

1) What is a particular university’s ‘geo-based’ impact (i.e., geographical reach) in the field of computer science as measured by the citations it receives from a variety of international universities?

2) Do high-ranking universities in the field of computer science tend to receive a broader geographical reach of citations than those that achieve a lower rank?

Our third research question relies also on Leiden and QS World University Ranking data, but includes: a) only conference proceeding publications matched to ranked universities, and b) data from the Computing Research and Education Conference (CORE) Ranking (2014).

3) Do computer scientists working at top ranked universities tend to participate more often in the top ranked CORE conferences than those from lower ranking universities?

Methods

Rankings, articles and proceedings data collection:

The Leiden Ranking data, the QS World University Rankings in the field of computer science, and the CORE Computing Research and Education Conference data are publicly available on the Web. By using a standard web scraping method, we will collect data from the following sites:

1) Leiden Ranking: http://www.leidenranking.com/ranking/2016/list

With the Leiden ranked list we will limit the universities to those associated with the field of ‘mathematics and computer science’ and use the basic P indicator (i.e., total number of publications), and percentile-based indicators: \( P_{top-10\%} \) and \( PP_{top-10\%} \). A percentile-based indicator is that which values publications based on their position within the citation
distribution of their field (Waltman & Schreiber, 2012). The $P_{\text{top-10\%}}$ is more precisely defined as: “the number of publications which belong to the top 10\% most frequently cited publications; a publication belongs to the top 10\% most frequently cited if it is cited more than 90\% if the publications published in the same subject area and in the same year” (Bornmann & Williams, 2013)

The indicator we use from the QS 2016 World University Rankings in Computer Science & Information Systems is computed as an “overall score” (a score of 0 to 100) for each university, and it is comprised of four components: 1) academic reputation (i.e., a global survey of academics), 2) employer reputation (i.e., a global employer survey), 3) research citations per paper, and 4) the h-index (see https://www.topuniversities.com/subject-rankings/methodology)

The Web of Science data for a set of computer science journal articles (doctype = Article,) for the period of 2007 to 2011 (publication years) has been provided to us from Clarivate Analytics (i.e. formerly Thomson Reuters). Our indicator for ‘geo-based’ impact is calculated in terms of citation variability, that is, the variability in the origin of citations received by the universities from other universities in countries worldwide (i.e., this is specifically for the articles published in 2007 to 2011, which have been cited during the period of 2012 to 2016). This approach, shown in the formula below, is adapted from what has been done earlier by Gao et al. (2013):

$$g_i(l) = \left| \bigcup_{i=1}^{P(l)} \bigcup_{j=1}^{C(l)} O(j) \right|$$

$I$: A certain institution

$P(l)$: Number of papers published by researchers affiliated with institution $I$

$C(i)$: Number of incoming citations for paper $i$

$O(j)$: Country of origin for citation $j$

And finally, a set of conference proceedings articles recorded in WoS for the year 2016 (doctype = Article,Proceedings Paper), will be matched to both their university of origin as well as the rank of the actual conference, as listed at the (CORE) Conference website. According to the CORE webpage: “conference rankings are determined by a mix of indicators, including citation rates, paper submission and acceptance rates, and the visibility and research track record of the key people hosting the conference and managing its technical program.” This includes a more detailed statement concerning the categorization ranks, which are labelled as A*, A, B, and C (see CORE, 2016).

For each university we will determine an overall CORE score based on the weighted proportion of its proceedings articles that match with one of the CORE rankings: A*, A, B, and C. Thus for each Core rank we will give an ‘A*’ a weight of 3, an ‘A’ a weight of 2, a ‘B’ a weight of 1 and a ‘C’ a weight of 0.5. If, for example, a university produces 5 articles each associated with the following CORE ranks (3=A*, 1=A, 1=B) we would obtain a CORE score = $[3(3/5) + 2(1/5) + 1(1/5)] / 5 = (1.8 + 0.4 + 0.2) / 5 = 0.48$. 

1133
Visualization:

The visualization component of this study is currently under development at the Department of Computer Science, University of Konstanz (see Figure 1, below). The aim is to provide users with an interactive tool that can support immediate geographical comparisons for a set of ranked universities from the field of computer science. A drop-down menu (left of screen) presents the university’s most recent rank, as per the Leiden ranking method for ‘mathematics and computer science’, the QS World Academic ranking in computer science, and associated Computing Research and Education Conference CORE score and geo-based impact.

The circular nodes on the map identify each university within its specific country. Red lines, or trajectories connect the university to other universities worldwide and illustrate the degree to which it is receiving citations. For instance, a user can click on a specific university node (University of Konstanz) in our geographical visualization tool, then hover over the trajectory and find a pop-up text at the right of the screen, which indicates total citations received for a specific time period. For example, (s)he might see a count of 30 incoming citations to University of Konstanz (Germany) from University of Toronto (Canada) for the period 2011-2016.

The countries in which the universities are situated will also be coloured as per their average geo-based impact. Again, the user can hover over a country, and at the right side of the screen a pop-up text will appear with the country name, its flag, and average geo-based impact, as calculated from the geo-based impacts of all of its regional universities’. A country with an above average geo-based impact will be coloured a darker shade of blue, and a country with a below average geo-based impact will be coloured a lighter shade of blue.

![Figure 1. Prototype of the geographical visualization tool for university rankings, CORE scores, and geo-based impacts in computer science.](image-url)

**Expected Results**

The results of this study will give scholars the opportunity to monitor their university’s current rank within a specific research field and to identify additional field-related measures
associated with these rankings. With the metric part of our study we expect to find that high-ranking universities in the field of computer science tend to receive a broader geographical reach of citations than those that achieve a lower rank. We further expect to find that computer scientists working at top ranked universities tend to participate more often in the top ranked CORE conferences, than those working at lower ranked universities. This will be shown on our geographical visualization tool in terms of multiple trajectories (red lines) corresponding to received international citations. It will also be shown on the basis of a high ‘geo-based’ impact measure and a high CORE score for each university in a pop-up menu beneath their corresponding Leiden and QS rankings.

References


Bornmann, L., de Moya Anegón, F. and Mutz, R. (2013), Do universities or research institutions with a specific subject profile have an advantage or a disadvantage in institutional rankings? *Journal of the American Society for Information Science and Technology, 64*, 2310–2316.


Using full-text data to create improved term maps

Nees Jan van Eck¹ Ludo Waltman¹ Min Song² Yoo Kyung Jeong²

¹{ecknjpvan, waltmanlr}@cwts.leidenuniv.nl
Centre for Science and Technology Studies, Leiden University, Leiden (The Netherlands)

²{min.song, yk.jeong}@yonsei.ac.kr
Department of Library and Information Science, Yonsei University, Seoul (Republic of Korea)

Abstract
A term map offers a visualization of a network of terms that co-occur in scientific publications. Term maps are usually created based on the titles and abstracts of publications. In this paper, we explore the use of full-text data for creating term maps. We create and compare a series of term maps based on the full text of publications in Journal of Informetrics. We use our results to discuss the advantages and disadvantages of different approaches for creating term maps.

Conference topic
Co-occurrence analysis; mapping and visualization

Introduction
Bibliometric analyses have traditionally been based on meta data of scientific publications. Increasingly, however, the full text of publications is becoming easily accessible for large-scale analyses. Some journals (e.g., journals published by Frontiers, PLOS, and Ubiquity Press and journals such as F1000Research and PeerJ) have websites from which the full text of a publication can be downloaded in XML format. This simplifies the large-scale processing of full-text data. For other journals (e.g., journals published by Elsevier, Springer, and Wiley), bulk downloaded of full-text data, in XML or some similar format, is possible through an API, provided that one has the required license.

The increasing availability of full-text data in convenient file formats offers lots of interesting possibilities for improving bibliometric analyses. In this paper, we explore one such possibility. We study the use of the full text of scientific publications for creating so-called term maps, and we compare the use of full-text data with a more traditional approach based on titles and abstracts. The term maps studied in this paper are created using VOSviewer (www.vosviewer.com; Van Eck & Waltman, 2010, 2014), a popular software tool for bibliometric visualization that is often used for creating term maps.

Term maps
A term map offers a visual representation of a term co-occurrence network. A term co-occurrence network is a network of terms in which the relatedness of terms is determined by their number of co-occurrences. Terms are often extracted from scientific publications, usually from titles and abstracts, and the number of co-occurrences of two terms typically equals the number of publications in which the terms occur together. A term co-occurrence network may for instance be constructed based on the publications that have appeared in a particular field of science.

To obtain an easy to interpret overview of the research that is done in a certain scientific field, it can be helpful to create a visualization of a term co-occurrence network constructed based on publications from that field. Such a visualization is referred to as a term co-occurrence map, or simply a term map. The VOSviewer software tool facilitates creating term maps. Example of term maps created using VOSviewer can be found in Figures 1, 2, and 3.
Interpreting a term map

A VOSviewer term map can be interpreted as follows:

- The size of a term reflects the number of occurrences of the term, for instance the number of publications in which the term occurs.
- The distance between two terms approximately reflects the relatedness of the terms, with strongly related terms typically being located close to each other. The relatedness of terms is determined based on their number of co-occurrences.
- The horizontal and vertical axes have no special meaning. A term map can be freely rotated and flipped. Since the distances between the terms remain the same, this does not affect the interpretation of the map.
- Colors are typically used to indicate clusters of strongly related terms. Terms that have the same color tend to be relatively strongly related to each other.
- There often is a considerable overlap of terms in a term map. When terms are overlapping, a label is shown only for the most frequently occurring term. For other less frequently occurring terms, only a circle is shown, not a label. When a term map is explored interactively using VOSviewer, it is possible to zoom in on a specific area in the map. This will reduce the overlap of terms, resulting in more labels being visible.

In the interpretation of a term map, it is recommended to focus on the general patterns emerging from the map. Focusing too much on individual terms is often less useful. A group of terms located close to each other, and typically having the same color, tends to represent a topic (or a research area) in a scientific field. The terms offer an indication of what the topic is about. By looking at the different groups of terms in a term map, one gets an impression of the main topics in a field. In addition, the distances between different groups of terms provide insight into the relatedness of the different topics in a field. Neighboring groups of terms tend to represent closely related topics.

Creating a term map

More information about the way in which term maps are created by VOSviewer is provided by Van Eck and Waltman (2011). For the purpose of this paper, the following points are of particular importance:

- Terms may consist of multiple words. Any sequence of nouns and adjectives, with the last word in the sequence being a noun, can be a term. Terms are extracted from publications, typically from titles and abstracts, using natural language processing algorithms. Plural terms are converted into singular ones.
- The terms included in a term map are selected using a term selection algorithm. The idea of this algorithm is to select terms that seem to relate specifically to one or a few topics in a field. More general terms are filtered out. This can be terms such as ‘conclusion’, ‘method’, and ‘result’, but also terms that, within a certain field, can be considered very general. For instance, in the field of oncology, the term ‘oncology’ is very general and is therefore likely to be filtered out. In addition to general terms, terms that occur in only a very limited number of publications are also usually excluded.
- Occurrences and co-occurrences of terms can be counted in two ways, either using a binary counting approach or using a full counting approach. In the binary counting approach, only the presence or absence of a term in a publication matters. The number of occurrences of a term in a publication does not matter. The full counting approach, on the other hand, does take into account the number of occurrences of a term in a publication.
To illustrate the distinction between binary and full counting, consider a publication in which terms A and B occur, respectively, two and three times. In the binary counting approach, this is counted as one occurrence of term A, one occurrence of term B, and one co-occurrence of the two terms. The full counting approach counts this as two occurrences of term A, three occurrences of term B, and six (i.e., two times three) co-occurrences of the two terms.

**Term maps based on full-text data**

When term maps are created based on the full text of scientific publications rather than based on titles and abstracts, a number of new possibilities emerge. The most straightforward way to count co-occurrences of terms is to count a co-occurrence for any pair of terms occurring in the full text of a publication. In this approach, if terms A and B both occur only once in the full text of a publication, term A in the first sentence and term B in the last sentence, this is counted as a co-occurrence. One could argue that in this situation terms A and B can hardly be considered to be related, and one may therefore prefer to count co-occurrences only for pairs of terms that are located sufficiently close to each other in the full text of a publication. This may mean that co-occurrences are counted only for pairs of terms that are located within a certain maximum distance from each other (i.e., co-occurrences are counted only within a certain window), where this maximum distance may for instance be expressed in terms of a certain number of words or characters. Alternatively, co-occurrences may be counted only for pairs of terms that are located within the same organizational unit in the full text of a publication, for instance in the same sentence, the same paragraph, or the same section.

We have explored a number of different approaches for counting co-occurrences in the full text of publications. In the rest of this paper, our focus will be on the approach in which co-occurrences are counted at the level of sentences, paragraphs, sections, or entire publications. We will contrast this approach with the traditional approach in which co-occurrences are counted in titles and abstracts.

**Data**

Our data set includes the full text of publications in *Journal of Informetrics* in the period 2007–2016. All publications of the Elsevier document types correspondence, editorial, erratum, full-length article, review article, and short communication were taken into account. The full text of the publications was downloaded in XML format using the Elsevier ScienceDirect Article Retrieval API. In total, the full text was obtained for 688 publications. Sections and paragraphs within the full text of a publication could be identified in a straightforward way using XML tags. In the case of sections, we focused on the main sections in a publication. If these main sections were subdivided into subsections, this subdivision was ignored. Sentences could not be identified using XML tags. Instead, we developed a sentence splitting algorithm to identify sentences. Table 1 offers some statistics for our data set.

**Table 1. Statistics for our data set of the full text of 688 publications in *Journal of Informetrics*.**

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Average per pub.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sections</td>
<td>4,126</td>
<td>6.0</td>
</tr>
<tr>
<td>Paragraphs</td>
<td>28,931</td>
<td>42.1</td>
</tr>
<tr>
<td>Sentences</td>
<td>131,443</td>
<td>191.1</td>
</tr>
</tbody>
</table>
Results

Creating the term maps

In total, we created ten term maps. Two maps were created based on titles and abstracts, one using binary counting and one using full counting. The other maps were created based on full-text data. In these maps, co-occurrences were counted at the level of sentences, paragraphs, sections, or entire publications and either binary counting or full counting was used.

As explained above, when creating a term map, general terms are typically left out. We left out 40% of the terms, which is the default choice in VOSviewer. We also needed to choose the minimum number of occurrences that a term is required to have in order to be included in a map. For each of the maps based on full-text data, we choose the minimum number of occurrences in such a way that the map includes about 2,000 terms. The maps based on titles and abstracts were created based on a smaller amount of data, and therefore we needed to follow a different approach for these maps. In the case of these maps, we used a quite low threshold for the minimum number of occurrences of a term. A term was required to have at least five occurrences.

The so-called attraction and repulsion parameters in VOSviewer, which determine the layout of the terms in a term map, were set to values of 1 and 0, respectively. These values generally yielded maps with a satisfactory layout. The resolution parameter, which determines the granularity of the clustering of terms, was set to its default value of 1. An exception was made for the three maps discussed in more detail in this paper (see below). For each of these maps, the value of the resolution parameter was fine-tuned in order to improve the consistency between the layout and the clustering of the terms.

Comparing the term maps

Due to space limitations, it is not possible in this paper to present a full comparison of all ten term maps. However, the maps have all been made available online. They can be explored interactively at [www.neesjanvaneck.nl/JOI_full_text_term_maps/](http://www.neesjanvaneck.nl/JOI_full_text_term_maps/).

We focus on comparing the following three term maps:

- Binary counting map based on titles and abstracts (Figure 1).
- Full counting map based on full-text data with co-occurrences counted at the level of entire publications (Figure 2).
- Full counting map based on full-text data with co-occurrences counted at the level of paragraphs (Figure 3).

In the case of the map based on titles and abstracts, we use binary counting because this is the default choice in VOSviewer and, consequently, because this is probably how term maps are typically created when VOSviewer is used. In the case of the maps based on full-text data, we use full counting because it turned out that full counting clearly yields more satisfactory results than binary counting, especially when co-occurrences are counted at the level of entire publications (see the maps available online).

Our findings resulting from comparing the three above-mentioned term maps can be summarized as follows:

- In terms of the level of detail that is provided, the maps based on full-text data are much richer than the map based on titles and abstracts. The latter map includes fewer than 400 terms, while the other two maps each include more than 2,000 terms. Especially when one zooms in on a specific area in a map, the full-text maps offer significantly more information than the map based on titles and abstracts.
- The full-text map based on paragraph-level co-occurrences has a more fine-grained structure than the full-text map based on publication-level co-occurrences. The former map for instance has an area in which many statistical terms are located closely...
together. Likewise, it includes areas showing the names of scientific fields and of countries. In the map based on publication-level co-occurrences, the structure is less fine-grained and similar areas are less well visible.

Which of the different term maps offers the most satisfactory representation of research published in *Journal of Informetrics*? In our view, there is no simple answer to this question. When a map is used as a static visualization, the increased level of detail provided by full-text data seems to be of limited value and a map based on titles and abstracts may offer sufficient information. On the other hand, when a map is used interactively, there seems to be a clear added value in the use of full-text data. One then has the possibility to zoom in on specific areas in a map to fully benefit from the detailed information extracted from the full text of publications.

When full-text data is used, the choice between the different levels at which co-occurrences can be counted is not entirely straightforward. Working at the paragraph level yields a more fine-grained structure than working at the publication level. This may be seen as an advantage. On the other hand, it seems that areas in a paragraph-level map do not always represent topics in the literature. For instance, an area showing all kinds of statistical terms does not represent a topic covered by *Journal of Informetrics*. There are not many publications in *Journal of Informetrics* that have statistical methodology as their core topic. Statistical terms are used in quite a lot of publications in the journal, but they usually do not represent the core topic of these publications. However, statistical terms probably tend to occur together in specific paragraphs in a publication, and this explains why a paragraph-level map includes an area in which many statistical terms are located closely together. This is an accurate reflection of the co-occurrence patterns of statistical terms at the level of paragraphs, but it does not represent a topic covered by *Journal of Informetrics*.

**Conclusion and future research**

In this paper, we have presented a first exploration of the possibility to create term maps based on the full text of scientific publications rather than based on titles and abstracts. Using the VOSviewer software tool, we have compared maps created in a number of different ways. Instead of identifying a single optimal approach for creating term maps, we have discussed advantages and disadvantages of different approaches.

There are a number of issues that can be addressed in future work. First of all, the analysis presented in this paper can be redone at a larger scale, for instance involving more than one journal. In addition, alternative approaches for counting co-occurrences of terms in the full text of publications can be explored. In particular, the idea of counting co-occurrences of terms that are located within a certain maximum distance from each other can be tested. Some practical issues also need further attention. For instance, term maps created based on full-text data include quite a lot of author names, and one may want to filter these out. Finally, the use of full-text data for creating other types of maps can be studied. Maps based on co-citation relations are especially likely to benefit from the availability of full-text data.

**References**


Figure 1. Binary counting term map based on titles and abstracts.

Figure 2. Full counting term map based on full-text data with co-occurrences counted at the level of entire publications. Some terms in the periphery of the map are not visible in the figure.

Figure 3. Full counting term map based on full-text data with co-occurrences counted at the level of paragraphs. Some terms in the periphery of the map are not visible in the figure.
Gender differences in research diversification behavior

Abstract

This work investigates whether and in what measure there occur gender differences in research diversification behavior by scientists. We analyze the nature of research diversification along three dimensions: extent of diversification, intensity of diversification, and degree of relatedness of topics in which researchers diversify. For this purpose we propose three bibliometric indicators, based on the disciplinary placement of scientific output of individual scientists. By classifying scientists by gender and field of research, we are able to assess gender differences along the above three dimensions in each field.

Keywords

Interdisciplinary research; diversity; specialization; relatedness; bibliometrics; university

Conference Topic

Research Fronts and Emerging Issues; Studies on the level of individual scientists
Introduction

Scientific research is becoming increasingly interdisciplinary (Porter & Rafols, 2009). This might well be the effect of a policy started in the mid-1990s by a number of funding agencies and higher education institutions, aimed at promoting the expansion of interdisciplinary research. At the same time, there were initiatives aimed at reducing the underrepresentation of women in science. The development of the interdisciplinary research (IDR) phenomenon has attracted the attention of many scholars and raised challenges on various fronts. Also gender differences in research have been the focus of many investigations, in particular female underrepresentation, performance differences and their determinants. Research has focused on each of these trends independently, but very few studies have considered their interaction, such as investigating whether there occur gender differences in the propensity to diversify one’s own research activity. Leahey (2006) contends that in addition to institutional issues, the extent of research specialization can also help explain the process by which gender affects research productivity. The preference of women for less specialised research, hinders their capacity to get published and cited. Unfortunately, her claim is based only on findings from two disciplines: sociology and linguistics. Rhoten and Pfirman (2007) showed that female scientists are much more likely than male scientists to engage in IDR projects. The authors consider also the expectation that women are well positioned to make major advances in interdisciplinary science. The current work intends to shed additional light on the issue of gendering of interdisciplinary science. It aims at assessing whether and to what extent there occur gender differences in field diversification of research activities. The analysis will be carried out at fine-grained field level, to detect possible differences across fields.

The investigation entails defining and measuring IDR. In the next Section we will review the literature on the subject. In Section 3 we present the methodology applied and the data used to conduct the analysis. In Section 4 we present the results, and in Section 5 the conclusions.

Literature review

A review of the literature on IDR studies shows that the issues receiving most attention are the taxonomic problem, the development of schemes for conceptual and practical definition of IDR, and subsequently its measurement (Huutoniemi, Klein, Bruun, & Hukkinen, 2010). According to a review by Klein (2008), the most commonly accepted scheme for definition of IDR is that involving three concepts: multidisciplinarity, interdisciplinarity and transdisciplinarity. Each of these is characterized by a particular type of “knowledge integration”, meaning the particular type of merging of theories and concepts, techniques and tools, information and data, from various fields of knowledge (Porter, Roessner, Cohen, & Perreault, 2006). According to Wagner et al. (2011) the phenomenon of knowledge integration can occur within a single mind, as well as within teams.

According to Stirling (1994), IDR displays some combination of only three basic properties, named “variety”, “balance” and “disparity”. Stirling (2007) also proposes the indicators suited to measurement of each: the “variety” indicator is defined as the answer to the question: “how many types of thing do we have?”; for “balance” it is instead the answer to “how much of each type of thing do we have?”; finally, “disparity” is “how different from each other are the types of thing that we have?”.

In the bibliometric sphere, these concepts have been widely applied in the investigation of IDR, as demonstrated in a review of the issue by Wagner et al. (2011). Many studies take a bottom-up approach, building from measurement of interdisciplinarity for individual articles.
The proposed measures are based on the disciplinary profile of the cited references, considering that reference to the preceding literature in various disciplines is as a signal of acquisition and integration of the results of these disciplines (Porter, Cohen, Roessner & Perreault, 2007; Rafols & Meyer, 2010; Wang, Thijs & Glänzel, 2015; Mugabushaka, Kyriakou & Papazoglou, 2016). In particular, Porter and Rafols (2009) used the works published in a cluster of selected journals indexed in the WoS over the period 2007-2011, examining their relative lists of references and identifying the disciplinary areas of the works cited, in terms of: i) number of subject categories (SCs) cited; ii) distribution of the citations among the SCs; iii) similarity or disparity among these SCs. In substance, Porter and Rafols (2009) proposed the measurement of Stirling’s (1994) three basic properties of research diversification through mapping the subject categories of cited publications. Zhang, Rousseau and Glänzel (2016) adopt the same approach to study the interdisciplinarity of journals.

Other studies are instead based on a top-down approach: again using typical disciplinary classifications such as WoS subject categories, they study the frequency distributions for scientific portfolios produced by defined units of analysis. For example, van Raan and van Leeuwen (2002) propose an approach for measurement of the IDR of a research organization through the percentage of its publications in each SC, or of citations received from each SC. Bourke and Butler (1998) had previously investigated the IDR conducted within Australian university departments, through the analysis of journals hosting the 1990-1994 publications authored by researchers of each department. Rinia, van Leeuwen, van Vuren and van Raan (2001) applied a similar approach in the Netherlands, analyzing the outcomes from a nationwide evaluation program of all academic groups in physics. As an aside, the intention of the last two works was to understand if IDR should be assessed in the same way as “disciplinary” research.

One of the areas of analysis little visited by scholars concerns IDR at the level of individual researcher. This is a strategic area, if one thinks of the challenges of complex research and the current recourse to policy aimed at fostering interdisciplinary work and thus influencing choices by the protagonists – the researchers themselves. The only contributions in the literature seems to be that from Schummer (2004), who carried out a coauthor analysis of nanotech journals in 2002-2003. By mapping “disciplinary” affiliation of coauthors, he was able to measure the IDR of each scientist in terms of interaction with the disciplines associated to all their coauthors. Using a similar approach, Abramo, D’Angelo and Di Costa (2012) analyzed the degree of collaboration among all Italian academics from different disciplines in order to identify the most frequent “cognitive combinations of knowledge” in research activity, drawing on 2004-2008 WoS publications by all Italian professors in the sciences. Abramo, D’Angelo, and Di Costa, F. (2017) studied diversification in research activities of Italian academics, along three dimensions: extent, intensity and relatedness of field diversification. We now intend to investigate gender issues in IDR. In other words, whether and in what measure female and male scientists tend to diversify their research activity, and if this tendency varies according to their belonging to different research fields.

**Data and methods**

**Methodology**

The results obtained from empirical investigations on research diversification depend on both the breadth of time for observation of scientific production and the schemes applied for sectoral classification of the scientific output and researchers. As for the first factor, the longer the observed time window, the wider the scientific portfolio of the subject under
investigation and, consequently, its alleged variety. As for the second factor, the outcomes of an analysis aiming at measuring diversification depends heavily on the classification scheme for research output and its grain. We take a top-down approach, analyzing the disciplinary placement of scientific output of individual scientists. We resort to Web of Science (WoS) subject category (SC) classification for the following reason.

Wang and Waltman (2016) provide a comparative analysis of the accuracy of WoS and Scopus journal classification systems. The authors firmly conclude that the “Web of Science performs significantly better than Scopus in terms of the accuracy of its journal classification system”. Probing beyond the issue of greater or lesser precision in assigning a journal to a field, the problem certainly originates in the classification schemes of the two repositories. The WoS scheme comprises 252 SCs covering the sciences (178), social sciences (47), and arts & humanities (27).Ⅰ In Scopus, each journal is instead assigned to one of the 334 “All Science Journal Classification” (ASJC) codes.Ⅱ At first glance the greater granularity of the Scopus classification scheme would seem an added value over WoS. However, examining the two schemes closely, we have discovered that:

- in the WoS, in addition to “Multidisciplinary Sciences” there are another nine “undefined” categories (e.g. “Chemistry, Multidisciplinary”), while in Scopus there are 27 categories with even more critical levels of non-definition, of the kind “Medicine (all)” or “General”;
- 32.8% of all the journals indexed in Scopus have at least one undefined code (compared to 8.7% in WoS), 21.7% have an undefined code and no other code (against 8.3% in WoS), and 4.2% of journals have still more than one undefined code;
- journals indexed in Scopus are on average assigned 1.87 codes, compared to 1.58 under WoS.

Although an unambiguously “correct” classification system has yet to be identified, the above findings led us to opt for WoS field classification.

To assess different diversification behavior across fields, we recur to the Italian “scientific disciplinary sector” (SDS) classification scheme for academics. In the Italian academic system all professors are classified in one and only one field (named “scientific disciplinary sector”, or SDS, of which 370 in all), grouped into disciplines (named “university disciplinary areas”, UDAs, 14 in all).Ⅲ In this study we focus on the sciences, for which the WoS coverage of publications by Italian universities is satisfactory. The sciences consist of 192 SDSs grouped into nine UDAs.Ⅳ Unfortunately, apart from the Italian one, the only other large-scale system known to us is the Norwegian Research Personnel RegisterⅤ compiled by the Nordic Institute for Studies in Innovation, Research and Education (NIFU). The NIFU system classifies scientists in 58 scientific fields grouped in five main domains (Rørstad & Aksnes, 2015). If Scopus recognizes 334 ASJC codes and WoS 252 SCs, reason would have it that the field classification of scientists should be numerically comparable. For the above reasons we opted for the Italian SDS classification of academics.

The next step is the construction of the publication portfolio of each individual scientist, from which we can proceed to disciplinary classification of the works (full counting). In the case of Italian academics, we can apply an algorithm for disambiguation of author names, developed by D’Angelo, Giuffrida and Abramo (2011), which allows us to assign the publications indexed in WoS to their relative academic authors.

**Dataset**

The dataset for the analysis is the 2004-2008 scientific production achieved by Italian professors in the sciences. The choice of a publication window quite far in the past is in consideration of a planned follow-up study with the aim of assessing whether output in
diversified fields is more influential in terms of citations than output in the prevalent field of the author: a longer citation window assures more robust and reliable results. Data on academics are extracted from a database maintained at the central level by the Ministry of Education, University and Research, indexing the name, sex, academic rank, affiliation, and the SDS of each professor. Publication data are drawn under license from the WoS Core Collection.

The overall dataset consists of 31,101 productive professors. Table 1 shows their distribution by UDA and gender, as well as the relative scientific production for the five-year period under observation. In the investigated fields, the share of female professors is 29%, varying across UDAs, with a minimum in Industrial and information engineering (13.3%) and a maximum in Biology (48.2%). Female professors co-authored 35.6% of total publications in the dataset, while males 90.6%. Civil engineering and Industrial and information engineering are the two UDAs with the major gender gap in co-authorships. The analysis of single-authored papers reveals that the share of female individual performers is lower than male. The same holds true analyzing papers authored only by professors of the same gender.

Table 1: Dataset of the analysis

<table>
<thead>
<tr>
<th>UDA*</th>
<th>SDSs</th>
<th>Professors§</th>
<th>Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Total</td>
<td>Authored by F</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>2,685 (31.1%)</td>
<td>14,741</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>2,416 (17.3%)</td>
<td>24,004</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>3,101 (38.4%)</td>
<td>24,407</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>1,004 (25.3%)</td>
<td>4,701</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>4,729 (48.2%)</td>
<td>28,097</td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td>9,163 (26.8%)</td>
<td>56,533</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>2,504 (32.9%)</td>
<td>10,516</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>1,104 (15.9%)</td>
<td>4,435</td>
</tr>
<tr>
<td>9</td>
<td>42</td>
<td>4,395 (13.3%)</td>
<td>33,816</td>
</tr>
<tr>
<td>Total</td>
<td>192</td>
<td>31,101 (29.0%)</td>
<td>179,506†</td>
</tr>
</tbody>
</table>

* Mathematics and computer science; 2 - Physics; 3 - Chemistry; 4 - Earth sciences; 5 - Biology; 6 - Medicine; 7 - Agricultural and veterinary sciences; 8 - Civil engineering; 9 - Industrial and information engineering

§ With at least one publication in the 2004-2008 period. Female data among brackets
† The total is less than the sum of column data due to multiple counting of individual publications authored by professors of more than one UDA.

Indicators

The disciplines in which research activity is classified often overlap, and generally have quite weak boundaries – the fields within them even more so. The confines are anything but static, since continuous scientific progress contributes to variation in the scope of the fields, as well as the birth of new ones and disappearance of old ones. As already said, the analysis of IDR must in any case involve some predetermined reference classification: in our study we use the WoS classification system. We associate each publication in the database with only one topic. By topic we mean the SC of the hosting journal in the case of a mono-category journal, or the combination of WoS SCs when the publication is issued in a multi-category journal.

The authors can thus be divided into two classes: those who diversify, meaning their publications fall in more than one topic; those who do not diversify, meaning their
publications fall in a single topic. We refer to these classes as “diversified” and “specialized” authors. Obviously the distribution of the scholars between the two classes depends on the breadth of the publication window observed, the intensity of publication, as well as the classification scheme for disciplines. The object of our study is the diversified authors and their gender. For each author, we can first of all identify the dominant topic in which the individual publishes, meaning the most recurrent SC or SC combination in their publication portfolio. Similarly, we can identify the dominant discipline in which the topic falls, meaning the most recurrent discipline or discipline combination. We consider the case of Maria Rossi (Jane Doe in English), professor in the SDS FIS/03 (Physics of matter), who in the period of observation produced eight articles published in four different journals (Physical Review B, Physical Review E, Chemphyschem and Physical Review letters). Given the classification of these journals under the WoS system, we have the distribution illustrated in Table 2. The eight articles fall in four different topics, belonging to two different disciplines/combination of disciplines. The dominant topic is subject category UK (Physics, condensed matter), given that half of Rossi’s publications fall in this topic; while the dominant discipline is Physics. We can also observe cases of more than one dominant topic, which above all is more likely when the subject’s number of publications is low.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Discipline</th>
<th>No of publications</th>
<th>WoS_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK (Physics, condensed matter)</td>
<td>Physics</td>
<td>4</td>
<td>243195800122; 245330200070; 260574500061; 251986500011</td>
</tr>
<tr>
<td>UF+UR (Physics, fluids &amp; plasmas; Physics, mathematical)</td>
<td>Physics</td>
<td>2</td>
<td>228818200106; 242408800041</td>
</tr>
<tr>
<td>EI+UH (Chemistry, physical; Physics, atomic, molecular &amp; chemical)</td>
<td>Chemistry; Physics</td>
<td>1</td>
<td>231971100043</td>
</tr>
<tr>
<td>UI (Physics, multidisciplinary)</td>
<td>Physics</td>
<td>1</td>
<td>229700800052</td>
</tr>
</tbody>
</table>

We will investigate three dimensions of diversification of research by individual professors. The first is “extent of diversification (ED)” given by the number of topics covered in the person’s scientific portfolio. The second is intensity of diversification, meaning the proportion of the professor’s output falling outside of their sector of specialization – measured by the indicator “diversification ratio (DR)”, given by the ratio of papers falling in topics other than the dominant one to the total number of publications. The higher the value of ED and the closer DR is to one, the more the individual’s research activity is diversified. The opposite situation denotes a highly specialized professor. There can also be antithetical situations: i) a high ED value jointly with a low DR value would indicate that the subject is predominantly specialized but open to exploring new fields; ii) a low ED value with a DR value tending to one is quite unlikely, unless the scientific production is very low. The last dimension investigated is the cognitive relatedness of the topics studied by the academic. Measurement of this requires definition of a threshold of cognitive proximity. For this purpose we associate the individual WoS topics to the disciplines, meaning we can identify the topics as “related” if they fall within the same discipline. The indicator for this dimension is “relatedness ratio (RR)”, equal to the ratio of number of papers in the dominant discipline to total number of papers. An RR of 1 indicates that the professor, although diversifying, does not go beyond their own disciplinary area. It is likely that a statistician (whose sphere of research can range from statistics to economics, medicine, agriculture, etc.) would have a much lower degree of relatedness than a surgeon.

According to the above taxonomy, for Maria Rossi we observe:
- An ED of 4;
- A DR of 0.5, since half of the total publications fall outside his dominant topic
An RR of 7/8, since 7 of the 8 publications are associated with the dominant discipline (Physics).

Analysis

In this section we analyze the three main dimensions characterizing research diversification, by the above indicators ED, DR and RR. Given the different operating definition of the indicators, the three datasets relative to ED, DR, RR analyses are of different sizes:

- **ED**: 31,101 observations in all, i.e. all professors with at least one publication in the period examined;
- **DR**: 26,943 observations in all, i.e. all professors with at least one publication falling in topics other than the dominant one;
- **RR**: 23,007 observations in all, i.e. all professors with at least one publication falling in disciplines other than the dominant one.

**Extent of research diversification**

We start the analysis at field level. As an example we consider the BIO/10 (biochemistry), falling in the UDA Biology. In this SDS 863 professors published at least once (in WoS indexed journals) in the period under observation (Table 3). The gender distribution shows 433 male professors and 430 female. Males’ average ED (6.635) is higher than females’ (4.588), but it is also true that males publish on average (14.02 publications) more than females (8.72). Controlling for publication rates, the average ED per publication presents no gender differences (0.508 vs 0.505).

**Table 3: Average number of publications and extent of research diversification (ED) for Italian academics in BIO/10-biochemistry (2004-2008 production), by gender**

<table>
<thead>
<tr>
<th>Gender</th>
<th>Obs</th>
<th>Average ED</th>
<th>Average output</th>
<th>Average ED per publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>430</td>
<td>4.588</td>
<td>8.721</td>
<td>0.508</td>
</tr>
<tr>
<td>M</td>
<td>433</td>
<td>6.635</td>
<td>14.023</td>
<td>0.505</td>
</tr>
</tbody>
</table>

The graph of Figure 1 confirms the strong correlation between ED and publication rates (Pearson ρ = 0.872), and shows the very large concentration of males among outliers (13 out of 14 professors with over 50 publications were male).

Aiming at investigating whether and in what measure there occur gender differences in research diversification behavior, we therefore adopt two different models of analysis:

- In the first model (Model 1) we assess the relation between ED and gender, via a direct OLS regression;
- In the second model (Model 2) we add the independent variable: number of publications by each single professor.

Results are presented in Table 4 for each SDSs of UDA Biology. With regard to Model 1, the impact of gender on ED is significant in over two thirds of the SDSs (13 out of 19) and is positive, meaning that being male leads in general to higher ED. As far as Model 2 is concerned instead, $R^2$ notably raises thanks to the introduction of output as independent variable. Alongside gender’s significance wanes: it remains significant in only five SDSs, and in two of them (BIO/02-Systematic Botanics e BIO/06-Comparative Anatomy and Citology)
it is negative, meaning that being female leads to higher ED.

**Figure 1:** Dispersion diagram of the number of publications vs average extent of research diversification (ED) for Italian academics in BIO/10-biochemistry (2004-2008 production), by gender

![Dispersion diagram](image)

**Table 4: Regression analysis of ED vs gender, for SDS of Biology**

<table>
<thead>
<tr>
<th>SDS</th>
<th>Obs</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIO/01-General Botanics</td>
<td>91</td>
<td>0.045</td>
<td>0.047</td>
</tr>
<tr>
<td>BIO/02-Systematic Botanics</td>
<td>94</td>
<td>0.060</td>
<td>-0.140*</td>
</tr>
<tr>
<td>BIO/03-Environmental and Applied Botanics</td>
<td>112</td>
<td>0.184</td>
<td>0.001</td>
</tr>
<tr>
<td>BIO/04-Vegetal Physiology</td>
<td>100</td>
<td>0.053</td>
<td>-0.039</td>
</tr>
<tr>
<td>BIO/05-Zoology</td>
<td>280</td>
<td>0.233**</td>
<td>0.025</td>
</tr>
<tr>
<td>BIO/06-Comparative Anatomy and Citology</td>
<td>256</td>
<td>0.019</td>
<td>-0.104***</td>
</tr>
<tr>
<td>BIO/07-Ecology</td>
<td>205</td>
<td>0.317***</td>
<td>0.059</td>
</tr>
<tr>
<td>BIO/08-Anthropology</td>
<td>52</td>
<td>0.163</td>
<td>0.166**</td>
</tr>
<tr>
<td>BIO/09-Physiology</td>
<td>590</td>
<td>0.174***</td>
<td>-0.007</td>
</tr>
<tr>
<td>BIO/10-Biochemistry</td>
<td>863</td>
<td>0.344***</td>
<td>0.026</td>
</tr>
<tr>
<td>BIO/11-Molecular Biology</td>
<td>207</td>
<td>0.366***</td>
<td>0.014</td>
</tr>
<tr>
<td>BIO/12-Clinical Biochemistry and Biology</td>
<td>162</td>
<td>0.562***</td>
<td>0.170**</td>
</tr>
<tr>
<td>BIO/13-Applied Biology</td>
<td>248</td>
<td>0.358***</td>
<td>0.041</td>
</tr>
<tr>
<td>BIO/14-Pharmacology</td>
<td>662</td>
<td>0.292***</td>
<td>0.031</td>
</tr>
<tr>
<td>BIO/15-Pharmaceutic Biology</td>
<td>76</td>
<td>0.329*</td>
<td>-0.026</td>
</tr>
<tr>
<td>BIO/16-Human Anatomy</td>
<td>318</td>
<td>0.316***</td>
<td>0.041</td>
</tr>
<tr>
<td>BIO/17-Histology</td>
<td>156</td>
<td>0.332**</td>
<td>0.066</td>
</tr>
<tr>
<td>BIO/18-Genetics</td>
<td>164</td>
<td>0.335***</td>
<td>0.059</td>
</tr>
<tr>
<td>BIO/19-General Microbiology</td>
<td>93</td>
<td>0.588***</td>
<td>0.278***</td>
</tr>
</tbody>
</table>

Dependent variable: Extent of diversification (ED). Independent variables: Model 1, “gender” (F=0, M=1); Model 2, “gender” (F=0, M=1) and total output. OLS estimation method. Statistical significance: *p-value <0.10, **p-value <0.05, ***p-value <0.01. "Obs" equals the number of professors with at least one publication in the period examined.

The same analysis has been repeated for all 192 SDSs in the sciences (Table 5). With regard to Model 1, gender has a significant impact on ED in 76 SDSs, being positive in 70 SDSs. At UDA level, it is positive in at least half of the relevant SDSs in Mathematics and
computer science (6 out of 10), Chemistry (7 of 12), Biology (13 of 19), and Medicine (26 of 52).

With regard to Model 2, significance drops to 33 SDSs and in 14 of them impact is negative, i.e. being female leads to higher ED.

### Table 5: Number of SDSs in each UDA with significant effect of gender on ED

<table>
<thead>
<tr>
<th>UDA</th>
<th>Total SDSs</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β &gt; 0</td>
<td>β &lt; 0</td>
<td>β &gt; 0</td>
</tr>
<tr>
<td>1-Mathematics and computer science</td>
<td>10</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>2-Physics</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3-Chemistry</td>
<td>12</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>4-Earth sciences</td>
<td>12</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5-Biology</td>
<td>19</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>6-Medicine</td>
<td>50</td>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>7-Agricultural and veterinary sciences</td>
<td>30</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>8-Civil engineering and architecture</td>
<td>9</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>9-Industrial and information engineering</td>
<td>42</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>All</td>
<td>192</td>
<td>70</td>
<td>6</td>
</tr>
</tbody>
</table>

**Dependent variable:** Extent of diversification (ED). **Independent variables:** Model 1, “gender” (F=0, M=1); Model 2, “gender” (F=0, M=1) and total output. OLS estimation method. Statistical significance: *p-value <0.10.

Finally, we have aggregated observations at UDA level (Table 6). According to Model 1 the impact of gender on ED results significant in each UDA and at overall level, with the only exceptions of Physics and Earth sciences. The positive regression coefficient indicates that males tend to diversify research activity more than females. Controlling for the different intensity of publication of males (Model 2), significance drops to two UDAs only, Mathematics and computer science and Biology; and in the first one are females that show higher ED.

### Table 6: Regression analysis of ED vs gender, by UDA

<table>
<thead>
<tr>
<th>UDA</th>
<th>Obs</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>R²</td>
<td>β</td>
</tr>
<tr>
<td>1-Mathematics and computer science</td>
<td>2,685</td>
<td>0.147***</td>
<td>-0.034*</td>
</tr>
<tr>
<td>2-Physics</td>
<td>2,416</td>
<td>0.058</td>
<td>0.001</td>
</tr>
<tr>
<td>3-Chemistry</td>
<td>3,101</td>
<td>0.187***</td>
<td>0.014</td>
</tr>
<tr>
<td>4-Earth sciences</td>
<td>1,004</td>
<td>0.045</td>
<td>0.001</td>
</tr>
<tr>
<td>5-Biology</td>
<td>4,729</td>
<td>0.270***</td>
<td>0.029</td>
</tr>
<tr>
<td>6-Medicine</td>
<td>9,163</td>
<td>0.217***</td>
<td>0.014</td>
</tr>
<tr>
<td>7-Agricultural and veterinary sciences</td>
<td>2,504</td>
<td>0.092***</td>
<td>0.004</td>
</tr>
<tr>
<td>8-Civil engineering and architecture</td>
<td>1,104</td>
<td>0.150***</td>
<td>0.005</td>
</tr>
<tr>
<td>9-Industrial and information engineering</td>
<td>4,395</td>
<td>0.072**</td>
<td>0.001</td>
</tr>
<tr>
<td>All</td>
<td>31,101</td>
<td>0.161***</td>
<td>0.009</td>
</tr>
</tbody>
</table>

**Dependent variable:** Extent of diversification (ED). **Independent variables:** Model 1, “gender” (F=0, M=1); Model 2, “gender” (F=0, M=1) and total output. OLS estimation method. Statistical significance: *p-value <0.10, **p-value <0.05, ***p-value <0.01. “Obs” equals the number of professors with at least one publication in the period examined.

**Intensity of research diversification**

The second dimension of diversification analyzed concerns the intensity of diversification, measured by the indicator “diversification ratio (DR)”, i.e. the ratio of the number of
publications falling in topics other than the dominant one to the total number of publications.

With reference to professors in BIO/10-biochemistry, gender differences by DR drop as compared to those by ED (presented in Table 3): males on average show a DR value (63.8%) similar to that of females (65.2%). Also the correlation between DR and publication rates decreases (Pearson ρ 0.298).

As expected ED e DR are correlated (Pearson ρ 0.573): as said before, the higher the value of ED and the closer DR is to one, the more the individual’s research activity is diversified or, viceversa, specialized. Figure 2 shows the dispersion of the two indicators regarding 814 professors in biochemistry (416 males and 398 females) with ED different from zero, subdivided into four quadrants drawn along the medians of the relevant distributions (5 for ED and 66.7% for DR). Occurrence of females in quadrant II (107 out of 179) is higher than males’. This means that the combination of below median ED and above median DR is a more a female trait. The opposite is true in quadrant IV (46 males out of 61). The chi-quadro test confirms the significance of such concentrations.

**Figure 2: Dispersion diagram of extent of research diversification (ED) vs diversification ratio (DR) for Italian academics in BIO/10-biochemistry (2004-2008 production)**

The OLS regression shows that gender impacts in a significant way the intensity of research diversification only in Biology e in Chemistry (Model 1 in Table 7). We observe that R² remains close to 0 even when output is included in the model as an independent variable. In Model 2, the regression coefficient is significant (and negative) in two UDAs only, Industrial and information engineering and Mathematics and computer science. The latter confirms what already observed along the dimension ED: it is females who show a higher propensity to research diversification in this discipline.
### Table 7: Regression analysis of DR vs gender, by UDA

<table>
<thead>
<tr>
<th>UDA</th>
<th>Obs</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\beta$</td>
<td>$R^2$</td>
<td>$\beta$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>1-Mathematics and computer science</td>
<td>2,143</td>
<td>-0.009</td>
<td>0.000</td>
<td>-0.023*</td>
<td>0.026</td>
</tr>
<tr>
<td>2-Physics</td>
<td>2,229</td>
<td>0.012</td>
<td>0.000</td>
<td>0.014</td>
<td>0.004</td>
</tr>
<tr>
<td>3-Chemistry</td>
<td>2,916</td>
<td>0.020*</td>
<td>0.001</td>
<td>0.000</td>
<td>0.050</td>
</tr>
<tr>
<td>4-Earth sciences</td>
<td>826</td>
<td>-0.007</td>
<td>0.000</td>
<td>-0.014</td>
<td>0.024</td>
</tr>
<tr>
<td>5-Biology</td>
<td>4,294</td>
<td>0.024***</td>
<td>0.002</td>
<td>-0.003</td>
<td>0.086</td>
</tr>
<tr>
<td>6-Medicine</td>
<td>7,744</td>
<td>0.012</td>
<td>0.000</td>
<td>0.003</td>
<td>0.008</td>
</tr>
<tr>
<td>7-Agricultural and veterinary sciences</td>
<td>2,034</td>
<td>0.008</td>
<td>0.000</td>
<td>-0.005</td>
<td>0.061</td>
</tr>
<tr>
<td>8-Civil engineering and architecture</td>
<td>847</td>
<td>0.012</td>
<td>0.000</td>
<td>-0.012</td>
<td>0.099</td>
</tr>
<tr>
<td>9-Industrial and information engineering</td>
<td>3,910</td>
<td>-0.013</td>
<td>0.000</td>
<td>-0.023**</td>
<td>0.078</td>
</tr>
<tr>
<td>All</td>
<td>26,943</td>
<td>0.010***</td>
<td>0.000</td>
<td>-0.001</td>
<td>0.024</td>
</tr>
</tbody>
</table>

Dependent variable: Diversification Ratio (DR). Independent variables: Model 1, “gender” (F=0, M=1); Model 2, “gender” (F=0, M=1) and total output. OLS estimation method. Statistical significance: *p-value <0.10, **p-value <0.05, ***p-value <0.01. “Obs” equals the number of professors with at least one publication falling in topics other than the dominant one.

### Degree of relatedness

We have finally analyzed the cognitive relatedness of the topics studied by academics who diversify their research. We have measured it by the indicator named “relatedness ratio (RR)”, given by the ratio of publications falling in disciplines other than the dominant one to the total number of publications. With reference to the SDS BIO/10-biochemistry, the distributions show that on average males (54.5%) present an RR slightly higher than females (53.1%). In this SDS, there occurs no correlation between RR and publication rates (Spearman $\rho$ -0.044).

The regression analysis at UDA level shows that gender impact is significant at overall level (last row of Table 8), as well in two UDAs, Physics and Biology (Model 2). In Model 1 the relation is significant also in Agricultural and veterinary sciences. Regression coefficients are positive in these UDAs, showing that males tend to diversify their research in fields more closely related than females.

### Table 8: Regression analysis of RR vs gender, by UDA

<table>
<thead>
<tr>
<th>UDA</th>
<th>Obs</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\beta$</td>
<td>$R^2$</td>
<td>$\beta$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>1-Mathematics and computer science</td>
<td>1,600</td>
<td>0.015</td>
<td>0.001</td>
<td>0.002</td>
<td>0.021</td>
</tr>
<tr>
<td>2-Physics</td>
<td>1,911</td>
<td>0.042**</td>
<td>0.003</td>
<td>0.032*</td>
<td>0.097</td>
</tr>
<tr>
<td>3-Chemistry</td>
<td>2,713</td>
<td>-0.003</td>
<td>0.000</td>
<td>0.009</td>
<td>0.012</td>
</tr>
<tr>
<td>4-Earth sciences</td>
<td>460</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.006</td>
<td>0.033</td>
</tr>
<tr>
<td>5-Biology</td>
<td>3,919</td>
<td>0.021**</td>
<td>0.001</td>
<td>0.031***</td>
<td>0.008</td>
</tr>
<tr>
<td>6-Medicine</td>
<td>6,784</td>
<td>0.003</td>
<td>0.000</td>
<td>-0.003</td>
<td>0.005</td>
</tr>
<tr>
<td>7-Agricultural and veterinary sciences</td>
<td>1,509</td>
<td>0.023*</td>
<td>0.002</td>
<td>0.021</td>
<td>0.005</td>
</tr>
<tr>
<td>8-Civil engineering and architecture</td>
<td>764</td>
<td>0.025</td>
<td>0.001</td>
<td>0.028</td>
<td>0.002</td>
</tr>
<tr>
<td>9-Industrial and information engineering</td>
<td>3,347</td>
<td>0.002</td>
<td>0.000</td>
<td>-0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>All</td>
<td>23,007</td>
<td>0.011***</td>
<td>0.000</td>
<td>0.007*</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Dependent variable: Relatedness Ratio (RR). Independent variables: Model 1, “gender” (F=0, M=1); Model 2, “gender” (F=0, M=1) and total output. OLS estimation method. Statistical significance: *p-value <0.10, **p-value <0.05, ***p-value <0.01. “Obs” equals the number of professors with at least one publication falling in disciplines other than the dominant one.
Conclusions

Since Cole and Zuckerman (1984) referred to gender differences in research productivity among academic scientists as a “puzzle”, scholars have widely investigated possible determinants of female underrepresentation and lower performance. Very few studies have considered whether there occur gender differences in the propensity to diversify one’s own research activity and if such difference may explain performance gap. In Leahey (2006) opinion the preference of women for less specialised research, hinders their capacity to get published and cited and can help explain the process by which gender affects research productivity. Unfortunately, such conclusions are based only on observations from sociology and linguistics. With a different but related perspective, Rhoten and Pbirman (2007) showed that female scientists are much more likely to engage in IDR projects than males. The current work has tried to shed additional light on this issue.

Based on observation of 2004-2008 scientific production of all Italian professors in the sciences, we demonstrated that gender impact on the extent of research diversification is meaningful both at overall level and at discipline level, with the only exceptions of Physics and Earth sciences. Males tend to diversify research activity more than females. However, this seems mainly due to the higher publication intensity of males: when controlling for individual output, significance drops to two disciplines only, Mathematics and computer science and Biology; and in the first one are females that show higher extent of research diversification, confirming both Leahey (2006) and the Rhoten and Pbirman (2007) views. Conversely, in Biology such view is contradicted.

The greater propensity to research diversification of females in Mathematics and computer science is confirmed along another dimension of diversification, the “intensity of diversification”, which points out higher intensity of diversification of women also in Industrial and information engineering.

The use of a third indicator measuring the relatedness of topics investigated by diversified professors allows us to state that males tend to diversify their research in fields more closely related than females both at overall level, as well in Physics and Biology.

We conclude that the attempt to solve the puzzle of gender differences in research is weak-willed: the different research diversification behavior of females with respect to males tends to vary based on the dimension of diversification analysed and the specificity of each discipline. Our findings suggest caution in identifying research diversification as a co-determinant of the gender productivity gap between males and females.

References


Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the
interdisciplinarity of journals. *Journal of the Association for Information Science and Technology*, 67(5), 1257-1265.

NOTES


iv Mathematics and computer sciences; Physics; Chemistry; Earth sciences; Biology; Medicine; Agricultural and veterinary sciences; Civil engineering; Industrial and information engineering.


vii In view of a future extension of the analysis to other personal determinants of diversification, we have excluded 2,682 productive professors whose age is not available.

viii Article, reviews, letters and conference proceedings

ix We have associated each WoS subject category to one and only one of the following disciplines: Mathematics; Physics; Chemistry; Earth and space sciences; Biology; Biomedical research; Clinical medicine; Psychology; Engineering; Economics; Law, political and social sciences; Multidisciplinary sciences; Art and humanities.

x Because the intensity of publication varies across the SDSs within each UDA (consequently, ED and output distributions are structurally different), we have rescaled each variable by the average of the relevant distributions excluding nil values (Abramo, Cicero, & D’Angelo, 2012)

xi In this case the number of observations drops to 747 (368 females and 379 males), since we observe only authors with at least one publication falling in disciplines other than the dominant one.
Do Google Scholar and Web of Science reflect women’s and men’s scholarly impact differently? A comparison of U.S. researchers in sociology and economics

Jens Peter Andersen1 Mathias Wullum Nielsen2

1 jpa@ps.au.dk
Danish Centre for Studies in Research & Research Policy, Aarhus University, Aarhus (Denmark)

2 mwn@stanford.edu
Department of History, Stanford University. Stanford (United States)

Abstract
Several studies have demonstrated differences in Google Scholars’ and Web of Science's coverage of citing publications. In this paper, we examine whether citation data retrieved from Web of Science and Google Scholar reflect women’s and men’s scholarly impact differently. Our study is based on a sample of 200 randomly selected U.S. authors in economics- and sociology-related research areas. Our results illustrate noteworthy gender disparities in the per-paper citation rates across the two databases. In sociology, we find that women benefit more than men, when impact assessments are based on data from Google Scholar, while the opposite is the case in economics. If our results prove to be robust in a larger data-set, based on a more exhaustive matching of documents, they illustrate how the selection of data-sources can have consequences for how individual scholars are assessed, e.g. in tenure or grant review evaluations.

Conference Topic
Science policy and research assessment; Journals, databases and electronic publications; Data Accuracy and disambiguation; The application of informetrics on evaluation

Introduction
Advances in information technology has spawned a growing use of citation-indices in the assessment of individual researchers’ scholarly impact. Today, citation-counts inform tenure, hiring and funding decisions in many disciplines (Nielsen, n.d.; O’Connor & O’Hagan, 2016; Stephan, Veugelers, & Wang, 2017; Van den Brink, Fruytier, & Thunnissen, 2013; Weingart, 2005), and individual citation-performance has been shown to correlate positively with academic salary levels (Leahey, 2007). Citations are becoming essential building blocks for academic success and status in the academy; and given their increasing importance as sieves for allocating opportunities and rewards, the data sources used in citation analyses deserve careful consideration.

This paper investigates the potential gender consequences related to the use of different databases in the assessment of individual researchers’ scholarly impact. Based on a sample of 200 randomly selected U.S. authors in economics- and sociology-related research areas (100 women and 100 men), we examine whether citation data retrieved from Web of Science and Google Scholar reflect women’s and men’s scholarly impact differently. For authors in economics, we find that men benefit considerably more than women, when citation-scores are based on Google Scholar data as opposed to Web of Science data. In contrast, women are put at a relative advantage when Google Scholar is used as data source in sociology. The paper proceeds as follows: First, we briefly discuss differences between Web of Science and Google Scholar and specify our focus. Second, we describe methodology and data. Third, we present and discuss preliminary results, and conclude.
Framing the problem

During the last decade, Web of Science’s longstanding role as primary data-source in citation-analysis has been challenged by new data-providers such as Google Scholar and Scopus. Studies comparing Google Scholar, Scopus and Web of Science reveal large variations in content coverage dependent on discipline and time period (Falagas, Pitsouni, Malietzis, & Pappas, 2008; Neuhaus, Neuhaus, Asher, & Wrede, 2006). Existing comparisons focusing on the coverage of citing publications, however, typically find that Google Scholar captures more unique citations than Scopus and Web of Science, especially in the newer literature (Bosman, van Mourik, Rasch, Sieverts, & Verhoeff, 2006; Harzing & Alakangas, 2016; Kulkarni, Aziz, Shams, & Busse, 2009; Meho & Yang, 2006). This is not surprising given that Google Scholar includes many publication types not covered by the other databases (e.g. doctoral theses, conference proceedings, anthology articles, books, regional and online journals and publication outlets in other languages than English). In this study, we examine whether a shift from Web of Science’s “restrictive” citation index to Google Scholars’ more “inclusive” citation tracker, Harzing & Mijnhardt (2014) has any implications for the relative citation-performance of women and men. A recent study by Harzing & Alakangas (2016) found that disparities in the coverage of citing publications between Web of Science and Google Scholar were larger for the social sciences than for the natural and health sciences (computer science and engineering are exceptions). Thus, we limit our focus to the social sciences; specifically, authors in economics- and sociology-related research areas.

Methods & data

Data collection

Given our specific interest in analysing the potential gender consequences of database selection, rather than attempting to make general conclusions on women's and men's citation-performance, we harvested articles based on individual Google Scholar profiles. These were subsequently matched to corresponding Web of Science records. To identify Google Scholar profiles for researchers in economics- and sociology-related areas, we created a background population of authors having published in journals categorised under sociology and economics in Web of Science between 2014 and 2016. Author groupings created by Caron & Van Eck (2014) enabled us to obtain additional information, such as email and institutional affiliation for each author. A total of 8,565 economics scholars and 3,126 sociology scholars from the U.S. formed the basis of our selection process. These were classified as men or women based on their first name, using the Gender API1 tool. Gender API draws on social media data to provide statistical probabilities on whether any given first-name belongs to a man or a woman, while accounting for country variations. As certain names yielded inconclusive results, we limited our dataset to names that could be classified with at least 95% certainty. Gender statistics for the dataset are available in Table 1.

Table 1. Demographic data for background population.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Economics, n (%)</th>
<th>Sociology, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inconclusive</td>
<td>1,384 (16.2%)</td>
<td>480 (15.4)</td>
</tr>
<tr>
<td>Female</td>
<td>1,468 (17.1%)</td>
<td>1,046 (33.5%)</td>
</tr>
<tr>
<td>Male</td>
<td>5,713 (66.7%)</td>
<td>1,600 (51.1%)</td>
</tr>
</tbody>
</table>

1 www.gender-api.com
From this background population, candidates were drawn in alphabetical order (which for this particular case should be as good as any other kind of presumed randomness), and a search for their profile on Google Scholar was performed. The user interface for this procedure presented candidates interchanging between men and women and both fields, to ensure that both genders and fields were represented equally, despite differences in their distributions. We limited our final sample to 200 Google Scholar profiles: 50 for each gender in each field. We used keywords available for each Google-Scholar profile, or departmental affiliation(s) to ensure that the scholars were indeed active within either economics or sociology-related research. Using the R-package ‘scholar’ (Keirstead, 2015; R Core Team, 2015), full publication lists were downloaded, including bibliographic information and citation counts.

A very large proportion of references in Google Scholar were to books, conference proceedings, teaching material, trade journals and other grey literature not captured by Web of Science. To identify publications registered in both Web of Science and Google Scholar, we ran complete reference strings, including authors, title, journal and source information past the CrossRef API2, retrieving DOI information on all publications with a match score above 95. This procedure allowed us to retrieve DOIs from 3,376 out of 21,844 documents. Of these, only 1,743 could be matched to a Web of Science record. The remaining references were run through a series of regex searches to identify journal, volume, issue, publication year and page information. Matching these to Web of Science increased our total sample of matched records to 4,903. Citation data from both sources were collected up until the most recent time3.

**Evaluation methods**

Our analysis is based on a comparison of raw citations up until the present. Given our focus on the relative difference between a given paper's citation score in Google Scholar and Web of Science, and not the general citation-impact of the paper, we do not normalize for field. We also compute citation-ratios such that the ratio, \( R \), of a given paper is the number of citations received in Google Scholar divided by those received in Web of Science. The ratio thus expresses the relative citation advantage of Google Scholar, or the increased ratio of included sources. For raw citation comparisons, we report arithmetic means and confidence intervals computed using one-sample t-test in R. For the ratios, we choose instead to report the harmonic mean, which is better suited for this type of variable. To calculate confidence intervals for the harmonic means in R, we use bootstrapping (Canty & Ripley, 2015) and generate harmonic means of 10,000 resamplings of each group. The confidence intervals are found by selecting the values at the 2.5\textsuperscript{th} and 97.5\textsuperscript{th} percentile.

**Results**

Figure 1 presents the raw distribution of citations per paper in economics, as given by the two databases and grouped by gender. The corresponding distributions for sociology are shown in Figure 2. It is obvious from these illustrations that the citation advantage of Google Scholar is large in all cases, and that men in general receive more citations, but especially in Google Scholar. The overlap of confidence intervals for men and women should, however, be noted, despite the rather large differences in means.

---

2 https://github.com/CrossRef/rest-api-doc/blob/master/rest_api.md

3 retrieved April 28, 2017
Figure 1. Raw citation distributions for Web of Science (WoS) and Google Scholar (GS), grouped by database and gender. Limited to economics.

Figure 2. Raw citation distributions for Web of Science (WoS) and Google Scholar (GS), grouped by database and gender. Limited to sociology.

Figure 3 displays the ratios and bootstrapping percentile confidence intervals between citation scores. First of all, the figure reveals how sensitive the analysis is to outliers, implying that interpretations should be made with extreme care. However, the mean distributions indicate quite large differences for women and men in both economics and sociology, and given our use of harmonic means, which is the most conservative of the mean-calculations, this is a quite clear observation nonetheless.
Discussion
To our knowledge, this paper is the first to examine whether citation data from Web of Science and Google Scholar reflect women’s and men’s scholarly impact differently. Our results illustrate noteworthy gender disparities in the per-paper citation rates across the two databases. In sociology, we find that women benefit slightly more than men, when impact assessments are based on data from Google Scholar, while the opposite is true in economics.
If our results prove to be robust in a larger data-set, based on a more exhaustive matching of documents and a larger sample of Google Scholar profiles, they illustrate how the selection of data-sources in research evaluations can have consequences for how individual scholars are assessed, e.g. in tenure or grant review evaluations. Further, research-based comparisons of women's and men's citation performance (see e.g. Nosek et al., 2010) using Google Scholar data may prove to be incompatible, with equivalent studies (focusing on the same discipline) using Web of Science data.
In the present research, we have regarded all possible papers from each scholar, regarding them as “their” paper, regardless of the number of co-authors or first-authorship. This can potentially skew the results, however, as each paper is linked to a particular profile, we expect the potential effect of this factor to be smaller, than if the body of papers had been randomly selected. It is important to note that the research presented here is in progress, and the process for data collection and cleaning will be enhanced and expanded in future work. This should considerably strengthen the preliminary findings presented in this paper.

References


Meho, L. I., & Yang, K. (2006). A New Era in Citation and Bibliometric Analyses: Web of Science, Scopus, and Google Scholar.


On a trajectory towards parity: an historical analysis of gender in funding from the National Science Foundation

Cassidy R. Sugimoto¹, Nicolas Bérubé², and Vincent Larivière²

sugimoto@indiana.edu
Indiana University Bloomington (USA)

² nicolas.berube.3@umontreal.ca; vincent.lariviere@umontreal.ca
École de bibliothéconomie et des sciences de l’information, Université de Montréal, Montréal (Canada) and Observatoire des sciences et des Technologies (OST), Centre Interuniversitaire de recherche sur la science et la technologie (CIRST), Montréal (Canada)

Abstract
Human capacity for innovation is not well-utilized in a system that favors one gender over another. Despite decades of policies in gender equity in science, disparities persist along several dimensions. One dimension that has been relatively understudied is the role of funding on gender disparities in science. To address this gap, we provide an historical analysis (1981-2016) of funding, by gender, from the National Science Foundation in the United States. We find that the mean rate of funds received by women is higher than that of men (and has been so since the early 2000s). However, men control the dominant share of funding and fellowships. As a work -in-progress paper, we discuss the extraction and cleaning of the data, provide an initial analysis, and describe the ongoing analysis of these data. The results provide policy makers with a much needed historical lens on progress towards parity, identifying areas of success and areas which require additional efforts in order to fund women at rates equal to their participation in science.

Conference Topic
Science policy and research assessment; Funding; Gender; Science of science

Introduction
Innovation is critical to economic development, yet is hindered by gender disparities in the scientific workforce (Etzkowitz, Kemelgo, & Uzzi, 2000). Despite more than a decade of policies aimed at achieving gender parity (Hagmann, 1999), women produce proportionally and comparatively less than men—in terms of both scientific articles (Lariviere et al., 2015) and patents (Sugimoto et al., 2015)—and their work receives fewer citations (Lariviere et al., 2015). The matriculation of students in undergraduate programs is not the problem: the OECD (2012) reports more female entrants than men in all but four countries (Germany, Turkey, Korea, and Japan). Rather, disparities are seen further downstream (Newsome, 2012)—women drop out of academe at a higher rate than males or fail to advance, creating more opportunities for men to progress through the ranks. As a recent editorial in Nature (2013) noted, in the United States and Europe, “barely one-fifth of full professors are women,” and progress on gender equality in science “seems to have stalled”. The so-called “leaky pipeline” phenomenon has been the focus of a number of studies seeking to understand what contributes to the attrition of women (Morgan, Gelbgiser, & Weeden, 2013), though this metaphor has been under increased scrutiny as an explanatory mechanism (Miller & Wai, 2015).

One possible alternative explanation for gender disparities in science is in the allocation of resources. The largest amount of funding for academic research comes from the federal government (SEI, 2012), and studies have shown that gender serves as a predictor of grant attainment (Jagsi et al., 2011; Jagsi et al., 2009). It has been suggested that bias in grant attainment (such as racial discrimination (Anon., 2011) and other success is due to cumulative advantage (Merton, 1973), with successful advisors unconsciously choosing advisees who
reflect their own social characteristics, thus perpetuating educational elitism in scholarship (RAND, 2005). This is exacerbated by the fact that younger faculty are less likely to receive federal support than their more senior colleagues (RAND, 2005). Women fare poorly in disciplines where research is expensive (Duch et al., 2012), and male authorship is disproportionately associated with contributing resources, whereas female authorship is significantly more likely to be associated with labor roles such as performing the experiment (Macaluso et al., 2016). Studies which have sought to identify gender disparities in federal funding have been hampered by the lack of demographic data collected on applicants, despite Title IX compliance regulations (Mervis, 2015). Additionally, studies have sought to examine the relationship between funding and outcome metrics, with varied results. Few studies have employed advanced statistical methods and large-scale integrated data to examine the relationship between funding and outcome metrics, and fewer still have examined how this relationship is moderated by gender and country of funding.

This study addresses the lack of historical information on rates of funding by gender by means of a diachronic analysis of grants awarded by the National Science Foundation. As a research in progress, this work presents the methods of data extraction and cleaning of the data, initial descriptive results, and plans for future analyses.

Data and Methods

Funding data for this paper was retrieved from the National Science Foundation (NSF) list of awards per year, which can be downloaded in XML format from the NSF website: https://www.nsf.gov/awardsearch/download.jsp. Each project funded by the NSF since the 1950s is available, including the amount received, the title of the project and program through which funding was obtained, as well as the institution funded, among other variables. While data prior to 1981 do not include the name of the principle investigator (PI), data after 1981 includes this information (given name and family name), which allows for the assignment of a gender to the PI of the project. Hence, data presented in this paper cover the 1981-2016 period.

In order to assign gender to given names, we used an algorithm that mines Wikipedia pages of individuals who have a specific given name. More specifically, our gender sorting algorithm uses the Wikipedia API for Python, which retrieves gender information by counting key words in pages of persons with a specific given name. Four steps are used to assign gender to a given name. First, all Wikipedia pages that starts with the given name followed by a space and an uppercase character are isolated. Those pages are normally about a person with the studied given name. Second, the summary text before the first section is analysed for all words between spaces, apostrophes, periods, commas and parentheses. If the sum of the number of occurrences of “he” and “his” in the summary is equal to or more than three times the sum of “she” and “her”, the page is identified as masculine. If the sum of “she” and “her” is equal to or more than three times the sum of “he” and “his”, the page is identified as feminine. If neither cases happen, the page is skipped. Third, once the page list is exhausted or 20 pages have been identified, the query is stopped. Finally, if equal to or more than three quarters of all identified pages are masculine, the given name is set as M. Likewise, if equal to or more than three quarters of all identified pages are feminine, the given name is set as F. The reliability of the algorithm has been tested by comparing it to the Gender.c package—a free name gender database of 36,940 identified. Testing showed gender agreement 99% of the time for these names.

Table 1 presents, for the 1981-2016 data, the number and percentage of awards and of their amounts. It shows that our method allowed for the assignment of close to 93% of awards, and of slightly less than 90% of their amount.
Table 1. Number and percentage of awards and of funding received, by gender assigned, 1981-2016

<table>
<thead>
<tr>
<th>Gender assigned</th>
<th>Awards</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>Female</td>
<td>64,624</td>
<td>17.5%</td>
</tr>
<tr>
<td>Male</td>
<td>277,055</td>
<td>75.2%</td>
</tr>
<tr>
<td>Initials</td>
<td>2,663</td>
<td>0.7%</td>
</tr>
<tr>
<td>Unisex</td>
<td>10,937</td>
<td>3.0%</td>
</tr>
<tr>
<td>Unknown</td>
<td>13,003</td>
<td>3.5%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>368,282</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Results

Figure 1 presents, for standard and continuing grants—which account for more than 71% of all funding awarded by the NSF over the 1981-2016 period—the mean amount of award obtained by women and men (left panel), as well as the proportion of total funding (right panel). It shows that, until the end of the 1990s, women’s grants were sizably smaller than those of men, with women obtaining percentages of men’s grants that oscillated between 62% and 94%. However, from 2001 onwards, women obtain mean grants that are slightly above those of men. In terms of proportion of total funding, we observe that, although men still account for a larger share of the funding—which is not surprising given their higher proportion of the workforce—their share is significantly decreasing, from 95% in 1981 to 73% in 2016.

In terms of fellowships awarded—which includes both graduate and postdoctoral—we also observe that, while men still account for the majority of funded projects, women share is increasing (Figure 2). More specifically, while women’s share of fellowships was less than 10% in the early 1980s, they account for more than 40% in 2016.
Limitation
The lack of disambiguation at the individual level makes per capita analyses difficult. This is one of the limitations of the work-in-progress that will be addressed in subsequent research. Furthermore, the name gender algorithm was only tested against a small set of names. A more robust analysis is necessary in order to adequately assess the rigor of this novel method.

Discussion and Future Research
Despite considerable concerns about rates of funding for women in science, there is scant data at the federal level on rates of funding by gender. The main federal provider of academic research funding in the United States is the Department of Health and Human Services (of which the National Institute for Health (NIH) is the primary funding body) (SEI, 2016). Funding provided by the NIH during the period 1997-2003 was studied by Jagsi et al. (2009; 2011). These studies found that women were awarded a third of K08 and 44% of K23 (both early career research awards) funded by the National Institute of Health (NIH) between 1997-2003. Using these data and follow-up data from a sample of 211 women, they found that women were less likely to subsequently receive an R01 (a major research project grant) and published, on average, fewer papers. Studies of funding from the National Science Foundation, which represented (as of 2013) more than 13% of academic research funding—second only to the NIH—has not been analyzed. Our work, therefore, expands both in agency and breadth our understanding of the relationship between gender and funding in science.

The mean amount of grant provided to women is slightly higher than that of men’s and has been since the early 2000s. However, the proportion of total funding remains higher for men than for women, though it has been steadily drawing towards parity since the beginning of the analysis (1981). In 2016, 73% of funding was distributed to men. A similar pattern can be observed for fellowship funding, which has decreased to less than 60% in 2016. However, both of these amounts are disproportionate with many other benchmarks: for example, men comprised 70% of fractionalized authorship on research publications from 2008-2012 (Lariviere et al., 2013) and earned less than half of the doctoral degrees awarded in the US in 2014 (NSF, 2017). This suggests that, despite a trajectory towards parity, funding amounts do not match the gender
composition of the scientific workforce. Of course, there are strong disciplinary differences in both funding and gender composition of the scientific workforce; therefore, further research must take discipline into account. This will require creating a disciplinary taxonomy from the NSF data, given that it is provided according to program.

Grants are intended as inputs in the scientific system. Therefore, it is useful from an indicator perspective to measure the relationship between this input and subsequent output. In future research, we will match the funding data within NSF will other productivity and impact measures (using a disambiguated data set from WoS). This will provide an indicator not only of the rates of funding by gender, but the result of this funding, accounting for differences in discipline and institution. These data will provide a comprehensive and historical understanding of the relationship between funding and gender in the United States for the National Science Foundation. Future research will extend these analyses to other agencies in the United States, such as the National Institute for Health, as well as to other nations.

Science has an indisputable impact on the economy and wellbeing of society. Progress is hindered when a given population is not given adequate representation in science or the resources to perform their activities. The present study provides policy makers with a much needed historical lens on progress towards parity in science funding, demonstrating where parity has been reached (i.e., in mean grant award) and where there is still work to be done.

Acknowledgments
This work was funded by the Science of Science Innovation and Policy program of the National Science Foundation in the United States (NSF #1561299).

References


Women in the Shadow of Big Men: The Case of Canada Excellence Research Chairs

Gita Ghiasi1, Vincent Larivière2,4, Catherine Beaudry3,4

1 g_ghia@encs.concordia.ca
Concordia University, Montréal (Canada)

2 vincent.lariviere@umontreal.ca
Université de Montréal, (Canada)

3 catherine.beaudry@polymtl.ca
Polytechnique Montréal (Canada)

4 Centre Interuniversitaire de Recherche sur la Science et la Technologie (Canada)

Abstract
Canada Excellence Research Chairs program—an award worth up to ten million dollars over seven years to attract and support world-renowned researchers and their teams to establish research programs in Government of Canada’s science and technology priority areas at Canadian universities—has been one of the most controversial governmental funding allocations in Canada. One of the main criticisms to this program is the absence of clear selection and recruitment criteria, including promulgation of standards for inclusion and diversity, which have resulted in lack of representation of women among Chairs. The main purpose of this study is to shed light on gender differences in scientific production and impact of publications induced by Canada Excellence Research Chairs program and to examine co-authorship collaboration patterns that are formed as a result of introduction of this program. Findings reveal that when Chairs are listed as main investigators of the scientific work (either last or corresponding authors), female-led papers receive higher rate of citations and are published in journals with higher impact. Although citation impact of papers that include collaborations with women are the highest, more than 78% of researchers of each gender repeat their collaborations, with their male peers on authoring more than one papers. Last but not least, this study concludes that collaborations with women are fragile and are dependent on the presence of central male researchers. Therefore, contributions of women to high impact research is effective as long as they are under the shadow of more central, influential and popular men.

Keywords
Canada Excellence Research Chairs Program; Gender; Bibliometrics; Co-authorship Network Analysis

Conference Topic
Science policy and research assessment

Introduction
Occupational gender segregation—a pervasive phenomenon expressing different distribution of women and men across occupations and sectors—has long been of central concern to science and public policy (Bettio, Verashchagina, Mairhuber, & Kanjuo-Mrčela, 2009; Charles, 2003; Fortin & Huberman, 2002). It is commonly differentiated along two dimensions: horizontal and vertical. Horizontal segregation is the differences in concentration of women and men in certain professions or sectors, whereas vertical segregation refers to representation of men and women at different levels or positions of labour hierarchy in terms of income, prestige, and the like. The vertical dimension has dovetailed with well-known notions of “glass ceiling” (Hymowitz & Schellhardt, 1986) and “sticky floor” (Noble, 1992), where visible or invisible barriers prevent women from advancing into top levels of hierarchy, and intrinsic forces tend to keep women at the lowest levels of the hierarchy (Meulders, Plasman, Rigo, & O’Dorchai, 2010).
Academia is no exemption and is not immune to gender segregation. Women in science also hit the glass ceiling—a hindrance to those trying to progress to senior ranks in academia (Caprile, Addis, Castaño, & others, 2012). However, the “leaky pipeline” metaphor (Berryman, 1983) suggests that women tend to leave science long before they face the glass ceiling, bearing witness to the higher attrition of women from science as they advance through the pipeline of science (the educational/professional system training scientists).

In response to the growing concern on women’s progression and retention in academia, science and technology (S&T) policies have typically centered either on getting more women into science and engineering (S&E) fields, or on affirmative action programs to remedy gender disparities along the science pipeline (Buré, 2007). However, increases in quotas as a stand-alone measure is likely to be futile to fix the leaky pipeline and glass ceiling for women in science, unless reinforced by systematic socio-economic strategies that alleviate inequalities in other arenas of society in ways that support productivity, diversity and inclusion (Alexander & Jacobsen, 1999; Cheldelin & Eliatamby, 2011, p. 239; Etzkowitz, Kemelgor, & Uzzi, 2000).

This study tries to articulate these issues into Canadian scientific system, delving into Canada Excellence Research Chairs (CERC) program—an award worth up to ten million dollars over seven years to attract and support world-renowned researchers and their teams to establish research programs in Government of Canada’s S&T priority areas at Canadian universities. These awards are acknowledged to be most generous available globally.

The CERC program was launched in 2008, the successor to government of Canada’s economic plan known as “Advantage Canada”, a plan to improve Canada’s global competitiveness as a knowledge-based society by creating pool of talents and effectively translating research into innovation to provide solutions and address concerns primarily in four fields, namely (1) Environmental Sciences (2) Natural Resources (3) Health Sciences (4) Information and Communications Technologies (Science-Metrix, 2015). Advanced manufacturing, social inclusion and innovative society, and other areas to benefit Canada have been added to the government of Canada’s priority areas after the first competition (Science-Metrix, 2015).

This program has been one of the most controversial governmental funding allocations in Canada over the past decade. The primary criticisms were twofold: (1) the program’s lack of focus on support and retention of local talents, and (2) the absence of clear selection and recruitment criteria, including promulgation of standards for equal opportunities, inclusion and diversity (Ghoussoub, 2013; The Globe and Mail, 2010, 2014). The attention to the latter concern was raised after selection of Chair holders in the first CERC competition in 2010, among which no women were even shortlisted for nomination. This issue provided an immediate impetus to review CERC Gender Issues by the Ad Hoc Panel and the lack of representation of women among CERCs was further dubiously explained by underrepresentation of women in top academic positions, low participation of women in Canada’s top four priority areas, and women’s vulnerability to mobility stress and career change (Science-Metrix, 2015). Canada’s effort to allay this concern was limited to a more transparent recruitment process report from universities, and accommodation of rising stars (rather than established researchers) and number of CERC positions open to all areas of research into the CERC program (de la Giroday, 2013; Science-Metrix, 2015). These initiatives resulted in the selection of only one woman out of nine additional Chairs under the second competition concluded in 2015. As of total, only one of twenty seven appointed CERCs is a woman.

The extreme underrepresentation of women in CERC program reflect one of the consequences of the glass ceiling and leaky pipeline phenomena, upon which attainment of equity and

diversity practices are obviated in Canada’s largest investment in research. This study shares the assumption that appointment of more female CERCs alone does not necessarily lead to promote equitable access to the program for qualified researchers of each gender. For this purpose, this study looks into co-authorship collaborations that are initiated by CERC program and tries to answer one main question: What are the gender differences in research output and scientific contributions that are induced by CERC program? Therefore, this study applies bibliometric and social network analysis (SNA) techniques to provide a cross-gender analysis of scientific production and impact, and collaboration patterns of Canada Excellence Research Chairs.

**Methodology**

Bibliographic publication data is gathered from Web of Science (WoS) database for Canada Excellence Research Chairs, where Chairs are affiliated to the Canadian university awarded in years 2008-2015. Since the focus of CERC programme is on advancement and diversification of Canadian scientific community, collaborators of CERCs are defined as Canadian-affiliated researchers who have authored one or more papers with CERCs. CERC program is launched in 2008 and full given names of authors are available in the WoS from 2008 onwards. Therefore, gender is assigned to the given names of CERCs and their collaborators, using U.S. Census, WikiName, Wikipedia, France and Quebec lists (More details in (Larivière, Ni, Gingras, Cronin, & Sugimoto, 2013)). Among 27 CERCs, 21 CERCs are identified with at least one paper published with their awarded Canadian university. A total of 712 articles were extracted from the Web of Science database, along with their 750 distinct authors.

The quantitative analyses are grounded on bibliometric indicators of scientific output, using the number of scientific publications as an indicator of productivity of a researcher, the normalized citations as research impact, and normalized Impact Factor (IF) as a journal impact indicator. The normalized citation impact of a paper is calculated as the average yearly number of citations received by a paper divided by the average yearly number of citations to all the papers from the same year, in the same subject area and of the same document type (Glänzel & Schubert, 2005; Moed, De Bruin, & Van Leeuwen, 1995). The normalized journal IF is calculated similar to the citation impact where the IF of the journal associated to each paper is considered instead of number of citations.

The proportion of scientific production of each gender is defined as a fractional count of articles, according to which each author account for 1/x count of authorship where x is the number of co-authors of an article affiliated to a Canadian university to which a gender is assigned. In this paper, orders of authors in the byline is considered as a measure of contribution, due to the fact that fields of research of CERCs follow contribution-based authorship ordering as a common practice, based on which first author position is typically given to younger researchers with lower professional rank and last author position is assigned to the principal investigator with high rank (West, Jacquet, King, Correll, & Bergstrom, 2013). Articles with CERCs listed as a corresponding author are also identified to help analyse the contribution of researchers of each gender where CERCs are responsible for the research project as a whole, including acquisition of funding, oversight of a research process and production of the final manuscript (Yank & Rennie, 1999).

The collaboration patterns of CERCs are mapped using the co-authorship network analysis, in which nodes represent authors and two nodes are connected when two authors co-author a paper. Each link (edge) is categorized into whether it connects two female authors (F-F) link, or one female and one male author (F-M), or two male authors (M-M). Weight of each link represents the extent to which two authors who have already collaborated on a paper, repeat
their collaborations by co-authoring another paper together and is referred to as collaboration repetition rate in this research.

In this study, the listed network measures in Table 1 are deployed to characterize the CERC’s co-authorship network, assessing the role of authors of each gender. CERC’s network of co-authorship is further visualized using Gephi’s (Bastian, Heymann, & Jacomy, 2009) Force-Atlas 2 layout, in which the proximity of two nodes is defined by the weight of the edge that links the two nodes.

Table 1. Network measures and their properties

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree Centrality (Freeman, 1979)</td>
<td>A node’s total number of connection</td>
<td>How collaborative an author is.</td>
</tr>
<tr>
<td>Betweenness Centrality (Anthonisse, 1971; Freeman, 1979)</td>
<td>Number of times a node is on the shortest paths between other nodes</td>
<td>How much control an author has over information in the network.</td>
</tr>
<tr>
<td>Eigenvector Centrality (Newman, 2004)</td>
<td>A centrality measure that is proportional to the sum of centralities of those it is connected to</td>
<td>To what extent an author is connected to the most connected authors.</td>
</tr>
<tr>
<td>Closeness Centrality (Sabidussi, 1966)</td>
<td>Proximity of a node to the other nodes in the network</td>
<td>How easily an author can reach/spread information in the network</td>
</tr>
<tr>
<td>Clustering Coefficient (CC) (range 0 -1) (Holme, Min Park, Kim, &amp; Edling, 2007; Watts &amp; Strogatz, 1998)</td>
<td>Ratio of total number of links that could exist for an actor to the number of real existing links.</td>
<td>How well direct collaborators of one author are connected to one another if the author is removed from the network. The lower the CC, the more important an author is in the network.</td>
</tr>
</tbody>
</table>

Findings

21 Chairs are identified with at least one paper affiliated to their awarded university, noting that other 6 Chairs that are not included in our publication data analysis are selected as CERCs from 2015 onward. Therefore, they had no paper published with a Canadian affiliation in years 2008-2015. Among these 21 Chairs identified, one is female (who is involved in authoring 22 papers) and 20 are male researchers (who are responsible for authoring 690 total articles). 214 female and 515 male authors were involved in CERCs’ authorship collaborations, being referred to as CERC collaborators in this research.

Although share of female authorship for CERCs is very limited (4%), women account for 25% of authorship among CERC collaborators (Fig. 1). CERC publications with female collaborators involved, have higher or equal scientific impact as those authored with male collaborators (Fig. 2). When women are lead-authors (first authors), their papers receive lower number of citations although being published in journals with higher impact (Fig. 3). This has been associated to Matilda effect (Rossiter, 1993) in science in several studies (Ghiasi, Larivière, & Sugimoto, 2015; Larivière, 2014), according to which women’s contribution to science is played down and attributed to their male peers (which in this case is in terms of the citation rate expected to be received by publishing papers in journals with higher IF than journals in which male first-authored papers are published). However, when women are last authors their papers receive lower citation rate (Fig. 3).
When Chairs are listed as last authors (which is often held by principal investigator or supervisor of the research project) (Table 2), female first-authored papers receive higher citation impact and are published in journals with higher IF (Fig. 4). These differences are even more conspicuous when Chairs are corresponding authors (main investigators of the research project) (Table 2; Fig. 5).

Table 2. Number of first-authored papers and share of authorship of each gender when CERCs are last authors or corresponding authors

<table>
<thead>
<tr>
<th></th>
<th>CERCs as last authors</th>
<th></th>
<th>CERCs as corresponding authors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First-authored papers (#)</td>
<td>Authorship (%)</td>
<td>First-authored papers (#)</td>
</tr>
<tr>
<td>Female</td>
<td>48</td>
<td>25%</td>
<td>7</td>
</tr>
<tr>
<td>Male</td>
<td>156</td>
<td>75%</td>
<td>11</td>
</tr>
</tbody>
</table>

Co-authorship Network Analysis

Details on CERC co-authorship collaboration network can be found in Table 3. Degree centrality is higher for women, which shows that female researchers are more collaborative than male researchers involved in CERCs’ co-authorship collaborations. However, men are more productive and average clustering coefficient is lower for men in the network. This confirms that direct collaborators of a female author are better connected if she is removed from the network, highlighting important position of male researchers within the network.
Table 3. Network properties of CERCs’ co-authorship collaborations (2008-2015)

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nodes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CERC</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>CERC Collaborator</td>
<td>214</td>
<td>515</td>
</tr>
<tr>
<td>Total</td>
<td>215</td>
<td>535</td>
</tr>
<tr>
<td><strong>Productivity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CERC</td>
<td>22.00</td>
<td>34.50</td>
</tr>
<tr>
<td>CERC Collaborator</td>
<td>2.02</td>
<td>2.36</td>
</tr>
<tr>
<td>Total</td>
<td>2.11</td>
<td>3.56</td>
</tr>
<tr>
<td><strong>Degree</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CERC</td>
<td>28.00</td>
<td>35.50</td>
</tr>
<tr>
<td>CERC Collaborator</td>
<td>10.79</td>
<td>9.35</td>
</tr>
<tr>
<td>Total</td>
<td>10.87</td>
<td>10.33</td>
</tr>
<tr>
<td><strong>Clustering Coefficient</strong></td>
<td>0.27</td>
<td>0.21</td>
</tr>
<tr>
<td>CERC Collaborator</td>
<td>0.90</td>
<td>0.86</td>
</tr>
<tr>
<td>Total</td>
<td>0.89</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Share of F-F collaborations are the lowest in CERC authorship collaboration network and majority of collaborations are between two male researchers (Fig. 6), which is expected due to the low share of women involved. Moreover, M-M collaborations are shown to be the strongest ties among other collaboration types (Fig. 7). At individual level, the female CERC have repeated her collaborations with other female authors at a higher rate than her male authors. However, among 20 male CERCs, this trend is only observed for three Chairs. The findings also reveal that only 17% of female researchers and 22% of male researchers involved in more than one paper with CERCs, repeat their collaborations more with their female colleagues. Hence, it can be concluded that for both genders, repeating an authorship collaboration with a male researcher is preferred.

![Figure 6. Share of different collaboration type](image)

![Figure 7. Collaboration repetition rate (edge weight) per collaboration types](image)

Along these lines, average citation impact of papers forming F-F collaborations are the highest compared to those creating F-M and M-M collaboration (Fig. 8). Therefore, in the CERC co-authorship collaboration network, female authors are collaborating with other female authors on highly cited papers.

---

*This is not associated to the data size, 86 distinct papers are identified for F-F collaborations, 192 papers for F-M collaborations and 303 papers for M-M collaborations.*
Figure 8. Average citation and journal impact of papers forming different types of collaborations

Since other centrality measures such as betweenness, eigenvector and closeness cannot be applied to disconnected networks, these measures are compared for researchers of each gender in each CERC cluster (CERCs and their connections to their collaborators). Female collaborators of the female CERC are associated to higher betweenness, eigenvector and closeness centrality compared to male collaborators of the CERC. Only 6 of male CERCs are involved in clusters where women have higher betweenness centrality and 5 male CERCs are located in clusters where women have higher closeness centrality and 2 in cluster with higher female eigenvector centrality, noting that these overlap. These measures show the important position of men in the network in majority of CERC clusters, denoting vulnerability of F-F connections and their strong reliance on a male researchers.

Visualization of CERC’s collaboration network is shown in Fig. 9. In this network node size is based on degree centrality of an authors and node colours represent different roles of authors. Edges are differentiated based on the collaboration type. The visualization also highlights a presence of male collaborators with high degree centrality where higher shares of female authors and F-F collaborations (orange links) are present, highlighting the fragility of F-F links and their reliance on collaborations with men.

Figure 9. Visualization of co-authorship collaboration network of CERCs
Conclusions

The main purpose of this study is to shed light on gender differences in scientific production and impact of publications induced by Canada Excellence Research Chairs program and to examine co-authorship collaboration patterns that are formed as a result of the introduction of this program. This program is highly male dominated: women account for only 13% of total authorship engendered by CERC program. However, share of female authorship is higher (25%) among CERC collaborators. Although only one woman has been selected as CERC since 2010, female collaborators of CERCs have shown to be as influential as their male collaborators by being involved in papers with higher or equal scientific impact as those of their male peers. Matilda effect (Rossiter, 1993) might be present at the level of citations, based on which when women are lead authors, their scientific work receive lower recognition (citation rate) than the rate that is expected to be received from the journals their papers are published in (Ghiasi et al., 2015; Larivière, 2014). When Chairs are listed as main investigators of the scientific work (either last or corresponding authors), female-led papers receive higher rate of citations and are published in journals with higher impact. This finding reveals that women make the largest contribution to high-impact publications of Chair-supervised projects.

Although average citation impact of papers that are forming collaborations with women (which include F-F or F-M collaborations) are the highest, more than 78% of researchers of each gender repeat their collaborations, with their male peers on authoring more than one paper. Last but not least, based on the network visualization and the centrality measures, this study concludes that F-F collaborations are fragile and might be strongly dependent on the presence of central male researchers. Therefore, this can be said that contribution of women to high impact research is effective as long as they are under the shadow of more central, influential and popular men. The results of this study might benefit Canada’s limited S&T gender-related policies, with a specific focus on development and implementation of systematic strategies to increase participation of women in Chair programs as students, colleagues and professionals. The results of this study underpin the importance of collaboration with female researchers, which is often overlooked in retention and inclusion policies of women in S&T, and can provide a baseline to develop gender-responsive policies facilitating forming collaborations with women and favouring the participation of female researchers in CERC.

References


An extension of the Characteristics Scales and Scores: isolating ‘inter-institutional’ from ‘intra-institutional’ variance in highly skewed real population distributions

Viorel Proteasa¹  Mihai Păunescu² Adrian Miroiu³

¹viorel.proteasa@gmail.com
West University of Timișoara, (Romania)

²paunescu.mihai@gmail.com
The National University of Political Science and Public Administration (Bucharest, Romania)

³ad_miroiu@yahoo.com
The National University of Political Science and Public Administration (Bucharest, Romania)

Abstract
In the present article, we propose and analyze the consequences of using a specific aggregation rule for measures of individual scientific impact into collective level measures (institutions, departments) for real populations of academics in a specific science field. The point of departure of our argument is that current aggregations of collective unit scores, such as those employed by the influential international rankings, but also national evaluation policies is that these are prone to ‘the exploitation of the great by the small’ (Olson 2002, 35), that is the high influence exerted by the top ranked units in the evaluation exercise. We construct our research question from this observation, focusing on the issue of measuring inter-institutional differences in terms of scientific impact in skewed populations, where such differences are obscured by high intra-institutional variance. We extend the initial scope of the Characteristics Scales and Score (CSS), in order to address this question. We analyse scientific impact by two measures, total number of citations and a rational variant of Egghe’s g-index. Our main findings relate to the properties of the ‘fair’ category within the CSS: linearity and proportionality. Our main argument is that an aggregate level scientific impact, calculated as a population’s central tendency measures on the basis of individual scores is methodologically more relevant, and also politically more appropriate when calculated on the basis of "fair" category. We use data from a real population of scholars: the academics in medicine schools within the six health sciences universities of Romania. As such, our study addresses the general issues outlined in the research question, in a particular setting, touching upon what we consider an under-researched topic in scientometrics: that of the characteristics of real populations of academics and its links with policies.

Conference Topic
University policy and institutional rankings

Introduction
In the past decade, evaluation of both individuals’ and universities’ or departments’ performance has become common place for funding higher education (Sabic 2015; Vercruysse și Proteasa 2012), not only rankings (Hazelkorn 2011). Amongst the different activities put under scrutiny in this endeavour, university research features probably most prominently.
Research is generally evaluated by appraising scientific productivity (number of articles), scientific impact (as measured through the number of citations, $h$-index, $g$-index etc.) or both. Individual results are aggregated into collective level score at institutional or departmental level, using elementary mathematical operations, such as sums or averages, or more complex measures, such as the successive or second order $Hirsch$ index (Schubert 2007). These operations represent the mathematical formalization of one of the big issues in social sciences: the micro-macro aggregation.

The point of departure of our argument is that current influential rankings based on overall sample mean citations and/or other h-type indices are prone to ‘the exploitation of the great by the small’ (Olson 2002, 35). We construct our research question from this observation, focusing on the issue of measuring inter-institutional differences in terms of scientific impact in skewed populations, with a severe intra-institutional variance. We develop our argument from the analysis of a real distribution of bibliometric research outputs: the population of six medical schools in Romania. As such, our research can be seen as a study case, but we use the data we collected for a significantly more general purpose, that of discussing the merits of the Characteristics Scales and Scores (Glänzel 2007, 2010; Glänzel și Moed 2002; Schubert, Glänzel, și Braun 1987) in shedding light on the truncated median part of real populations of scholars. We end up proposing a measure of the central tendency, which incentivizes achievement along a wider share of academics, which are relevant for the research activities of academic departments. In the discussion section, we address some of the characteristics of real populations of academics in contrast to bibliometric populations, as well as the link between bibliometrics and policy.

**The problem and the research question**

D’Angelo and Abramo (2015) argue that the purpose of bibliometrics is to provide support for policy-making and decision-making in the achievement of institutional objectives, through the provision of methods and indicators for the evaluation of performance. However, when the evaluation refers to the performance of the collective level entity (department, institution) we enter a situation which has striking similitudes with one of the ‘classical’ theories in social sciences, the theory of collective action (Olson 2002). In short, Olson argues that if a group is large enough, individuals will have the tendency to act as free-riders i.e. to expect others to contribute while not-contributing. He advances some solutions: informal coordination, internal organizations, and selective benefits; at the same time, in small groups collective goods can be produced if few individuals pay the entire costs, “*a surprising tendency for the "exploitation" of the great by the small*” (Olson 2002, 35).

If we conceive the department score as a collective good, issues of production and consumption can be raised. In terms of consumption, the reputational benefits or the surplus or deficit budget associated with aggregate scores is distributed to the members of the department according to institutional frameworks and internal decisions. In terms of production, in the absence of effective informal coordination, internal organization or selective benefits, some department members may tend to behave like free-riders in Olson’s theory and exploit the greater contributors. We find this problem particularly relevant due to the skewness of scientific production and impact (Seglen 1992). We argue that the aggregation rules used by the indicators which produce distributional consequences are very important in such situations. They may either highlight, or obscure the contributions of some individuals or groups of individuals, by reducing these to a single collective level measure or index.

Let us illustrate this fact by briefly looking at how three influential rankings quantify research: Shanghai Jiaotong’s Academic Ranking of World Universities (ARWU)\(^ii\), Times Higher Education’s World University Rankings (THE)\(^iii\), and Quacquarelli Symonds’ World University Rankings (QS)\(^iv\). In terms of bibliometric measures, QS uses citations per faculty,
THE uses publications per faculty, while ARWU uses also publications, absolute counts and per faculty. All these measures are prone to „the “exploitation” of the great by the small” (Olson 2002, 35), due to the aggregation procedure: averaging, and summing – only in the case of ARWU. Without entering lengthy discussion, we will note that the other measures: reputation and research funding from competitive sources, are also prone to the same phenomenon, determined by different mechanisms.

Therefore we consider aggregation to be more than a methodological option and we will discuss it having in mind its political implications. We consider aggregation of collective level research scores to be an instance of a wider methodological problem in the social sciences: the micro-macro transition, or how do we aggregate, in a meaningful way, properties of the individuals in order to make valid inferences at collective (social system) level?

Due to the extreme skewness of individual scientific impact, we reformulate the problem as how can we measure the collective level inter-institutional variation, when the intra-institutional variation (differences in measures of individual scientific impact) has such a magnitude. A second, kindred methodological challenge, arises, namely that of dealing with the magnitude of top achievers’ impact. We use the Characteristic Scores and Scale (CSS hereafter) to account for the inter-institutional variance, with the aim to compare not entire populations, but rather truncated segments of the distributions. Thus, we further detail our research questions:

(1) Can we identify within the CSS an appropriate gauge of inter-institutional variance? We judge appropriateness both in methodological terms - a measure that is able to differentiate between various institutional performance levels - as well as in political terms, namely as relevance for a public policy aiming to raise the collective levels of scientific impact by providing incentives to larger segments of individuals

(2) Is that measure stable in the face of rare occurrences of very high citation?

The data we use comes from a real population of scholars: the academics in medicine schools within the six health sciences universities of Romania. As such, our study addresses the general issues outlined in the research question, in a particular setting.

Data and methods

The population we studied comprises of 3374 academics from the departments of medicine within the six health studies universities in Romania. We compiled the personnel lists from public sources: websites and reports. We established a comprehensive list of academics at department level, of their publications (including single and co-authored) and their respective citations from Scopus data base for a 5-year window between 2009 and 2014, excluding authors’ self-citations. The data were collected between November 2014 and May 2015. Full counting was used. We disambiguated manually authors’ name and their respective set of publications and, consequently, the corresponding citations to this set of publications, irrespective of the author’s affiliation at the time of the publication. We included in our research all the academics which we considered relevant for the medical sciences – persons teaching various medical, surgical and paraclinical subjects – according to an official categorization of the health studies in Romania, plus the fundamental sciences. These are the large majority of the academic staff in the medical schools in Romania. We also excluded a set of publications that not only were outliers in terms of citations, but also not actually belonging to the realm of research papers: guidelines, medical procedure recommendations, and definitions.

We analyse the distribution of total citations within the academics in our data base. As we are also interested in a particular policy context, the Romanian one, we add a second measure to our analysis. We prefer to use a rational variant of Egghe’s index instead of the measure of impact within the official funding methodology. The indicator within the methodology is a compound measure constructed on the basis of Hirsch index at 2/3 power as the individual
measure of impact, calculated for each of the three data bases, Web of Knowledge, Scopus and Google Scholar, and averages as aggregation rule. We adapted from (Tol 2008) the following variant of Egghe’s index:

\[ g_{\text{rational}} = g + \frac{\sum_{i=1}^{g+1} c_i - g^2}{(g + 1)^2 - g^2} \]

where \( g \) represents the value of Egghe’s index and, \( c_i \) represent the citations received by the \( i \)-th paper of the author, ordered descending after the number of citations.

This variant has the advantage of preserving the monotonicity properties of the G index and yet it introduces differences in the subunit interval. It does so by dividing the subunit in \([ (g + 1)^2 - g^2 - 1 ]\) equal intervals. Its specter of subunit values is discontinuous and increases with the value of \( g \). Its value is determined by the citations of the papers in the \( g \) core, plus the citations ascribed to the \( g + 1 \)th paper, where the papers are ranked decreasingly after their citations. For the population we study, its distribution is strongly correlated with that of the Hirsch index at 2/3 power: 0.91, Pearson correlation. The Hirsch index has the disadvantage of a discrete and rather limited range of values: for more than 99% of the authors the index takes values lower or equal to 10, while for almost half of the population (47.87%) the value is zero.

Findings

At a general descriptive level, the distributions we obtained by ordering decreasingly the academics according to their total citations can be fit with a fixed growth function (power law, but also reasonably an exponential or logarithmic, with one exception – the logarithmic fit of citations), which illustrates the skewness of the distributions. We added a unit to the entire distribution, in order to calculate logarithmic regression models – see (Thelwall 2016) for a detailed discussion on the method.

Table 1: Goodness of fit for the regression models, citations and \( g \)-index, the rational variant, for the entire population of academics

<table>
<thead>
<tr>
<th>Model summary, ( R^2 ), significance</th>
<th>Citations</th>
<th>Egghe’s G (rational)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logarithmic</td>
<td>.392***</td>
<td>.944***</td>
</tr>
<tr>
<td>Power law</td>
<td>.946***</td>
<td>.901***</td>
</tr>
<tr>
<td>Exponential</td>
<td>.925***</td>
<td>.973***</td>
</tr>
</tbody>
</table>

Other scholars use functions belonging to the same set to describe such distributions (Albarrán et al. 2011; Chatterjee, Ghosh, și Chakrabarti 2016; Perianes-Rodríguez și Ruiz-Castillo 2015; Radicchi, Fortunato, și Castellano 2008; Waltman, Jan van Eck, și Van Raan 2012). Similar curve estimations and comparable fit (\( R^2 \)) was obtained by (D’Angelo și Abramo 2015) for real populations of academics in various scientific disciplinary fields in Italy. Even three extreme outliers are not included in the citations’ model, the logarithmic fit increased considerably to a square \( R \) of .653, while the exponential and power law fit increase marginally.

We plotted each of the six the medical schools’ distributions on the same graph in Figure 1, below, also capping the upper end of the distributions for illustrative purposes. The advantage of analysing the two measures comparatively is that they vary in terms of skewness (internal variance). We note that the six distributions, each of them associated with a medical school, follow similar patterns to the populations. The distributions are shifted to the right of X axis, revealing differences in magnitude (number of academic staff) and also slightly upwards on the Y axis, revealing higher values of the bibliometric measures for some of the medical schools. One can also notice that the distributions tend to merge at the top, reflecting similarities as
regards the top achievers. They also converge towards the same minimum, but have different
tails reflecting differences in size. The graphic illustration cannot provide an answer to the
question: are there significant differences between the six distributions or they differ only in
terms of magnitude?

Figure 1: Distributions of citations and values of Egghe’s G index, the rational variant

Since the data are skewed and do not match a normal distribution, we test the null hypothesis
of no significant differences of scientific impact between the six departmental groupings of
academics with a Kruskal-Wallis test. The Kruskal-Wallis test is a chi-square distribution of
the sum of the squares of the group mean ranks. The null hypothesis is rejected at .000
significance level, thus we conclude that the six distributions differ not only in terms of size
different N), but also in terms of scientific impact, on both citations and g-index.

Once we established that the distributions exhibit differences which can be attributed to
differences in collective level scientific impact, we can further our analysis in order to answer
the research question of this article: how can we measure the inter-institutional variation, when
the intra-institutional variation has such a magnitude?

We proposed to use CSS to account for the inter-institutional variance, with the aim to compare
not entire populations, but rather truncated and thus comparable segments of the distributions.
In the original article, the authors split a population of articles based on their citedness into five
segments: uncited, poorly cited, fairly cited, remarkably cited and outstandingly cited. The first
cut-off point is determined by computing the average of the non-zero distribution, the second
cut-off point is determined by computing the average of the distribution above the first cut-off
point and so on. The authors use the categories to compare journals on the basis of the
proportion of articles in each category, mentioning that a similar approach can be used also for
department comparisons. The CSS had been used by (Albarrán et al. 2011) and (Albarrán, Perianes-Rodríguez, și Ruiz-Castillo 2015) for comparisons of distributions across 219 sub-fields, respectively 36 countries, (Ruiz-Castillo și Costas 2014) for comparisons distributions over 30 broad scientific fields, and (Perianes-Rodríguez și Ruiz-Castillo 2015) for comparisons of distributions across 500 universities in the Leiden ranking. Generally the CSS is used in relation to citedness, but, as the distributions of the two impact metrics have a reasonably similar shape, we extend the use of the CSS to Egghe’s index, the rational variant.

We thus use the characteristic scores and scales for the analysis of both citations distribution and rational g-index distribution as a way to truncate and meaningfully compare sub-samples of the entire populations, relevant for our policy evaluation purposes. As Glänzel (2010) argues, both h-type indices related indicators and characteristic scores proved useful for truncating the ranked sample and/or for selecting ‘top publications’. Furthermore, CSS can also be used for ranking various entities (journals) according to the amount of 'highly' cited papers they publish. Alternatively, we are instead interested in selecting 'fair' type units as we consider those more appropriate for our “real populations” departmental evaluation purposes and also, for the reasons explained above, opt for the rational g variant instead of the h-index.

One of the main methodological reasons why we adopt such a strategy is the fact that the six distributions are commensurable up to at least the 95th percentile, when extreme variation starts to occur (see Error! Reference source not found.). The intervals between the 50th and the 95th percentile are fairly stable, the universities' distributions grow linearly having comparable slopes.

Figure 2: Percentile analysis

Such a behaviour of the data is similar to the findings of previous articles that studied citations distributions across scientific fields (Crespo, Li, și Ruiz-Castillo 2013), respectively across universities (Perianes-Rodríguez și Ruiz-Castillo 2015).

A measure or inter-institutional variance based on the CSS

We adjusted the categories defined in the original version of the CSS in order to fit our purpose of comparing the six distributions. Due to the size of our population, we limited the CSS categories to three: (1) the “low” contributors – the academics with a value of the metric less or equal to the average; (2) the “fair” contributors – the academics with a value of the metric which is greater than the average and lower than the second average; (3) the “high” contributors - the academics with a value of the metric higher than the second average. A similar approach was adopted by (Ruiz-Castillo și Costas 2014). Unlike the original version of the scale, we included in the „low” category the academics with zero impact as well. They constitute the bulk
of the population of the medical schools and we presume that their percentage in the departments reflects the interplay between national policy and internal organization. We used Chi-square to test whether the two averages divide the six populations in quasi-similar proportions in all six medical schools for the two measures taken into consideration. The null hypothesis is rejected both for both impact measures, thus the six institutional distributions of the both measures are fairly similar and stable. This is in line with the conclusion of Glänzel (2007), who shows the class proportions or the characteristic scales are fairly stable across time and across fields.

Table 2: Chi square tests for CSS proportions

<table>
<thead>
<tr>
<th></th>
<th>Citations</th>
<th>G-index (rational)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>2.47</td>
<td>5.14</td>
</tr>
<tr>
<td>df</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Sig.</td>
<td>0.99</td>
<td>0.88</td>
</tr>
</tbody>
</table>

The 'fair' category of the CSS falls roughly within the 50 – 95 percentiles discussed in the percentile analysis in Figure 2. Glänzel (2007) notes that the class sizes of the individual subfields may somewhat differ from the generally applied rule of thumb (75–18–5–2%), but the deviations from this 'standard' are not dramatic. The variation among the scores is nonetheless considerable. We checked for the linearity of the distributions within the 'fair' category against that of the 'high' category, by comparing the determination coefficients $R^2$ to assess the goodness of a linear function fit. The results are presented in Table 3, below:

Table 3: Linearity of the 'fair' categories

<table>
<thead>
<tr>
<th>University</th>
<th>Citations 'fair'</th>
<th>Egghe's g (rational) 'fair'</th>
<th>Citations 'high'</th>
<th>Egghe's g (rational) 'high'</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>Slope</td>
<td>$R^2$</td>
<td>Slope</td>
</tr>
<tr>
<td>Bucuresti</td>
<td>.900</td>
<td>-.445</td>
<td>.325</td>
<td>.977</td>
</tr>
<tr>
<td>Cluj</td>
<td>.937</td>
<td>-1.086</td>
<td>.256</td>
<td>.967</td>
</tr>
<tr>
<td>Iasi</td>
<td>.906</td>
<td>-.763</td>
<td>.373</td>
<td>.923</td>
</tr>
<tr>
<td>Timisoara</td>
<td>.913</td>
<td>-.759</td>
<td>.818</td>
<td>.940</td>
</tr>
<tr>
<td>Craiova</td>
<td>.876</td>
<td>-1.060</td>
<td>.782</td>
<td>.961</td>
</tr>
<tr>
<td>Tg. Mures</td>
<td>.943</td>
<td>-.401</td>
<td>.640</td>
<td>.933</td>
</tr>
</tbody>
</table>

As it can easily be seen, the "fair" categories in both "raw citations" and "Egghes's g-index rational" distributions approximate very well linear functions (high $R^2$ coefficients for the linear fit), whereas the "high" categories' values are considerably less linear. In Figure 3, below, we plot the 'fair' category distributions within the six populations for illustrative purposes. A shift to the right reflects differences in size, whereas a vertical shift reflect differences in citation impact.
Figure 3: The 'fair' categories of the six medical schools

By using the CSS to truncate the distribution and select the ‘fair’ partition, not only we translated the research question from the analysis of a power law function to a linear one – which consequently makes comparisons amenable to parametric testing, but we also controlled for much of the distributional variation inside the medical schools. The segment represented by the 'fair' category behaves approximately similar in all six cases. Since the characteristic scales are similar, we should thus be able to control the intra-institutional variance of scores in order to compare those across the institutions in a meaningful way. We assessed the capacity of the CSS ‘fair’ partition to be used for comparing departments by carrying out a Kruskal-Wallis H test for citations as well as for the g-index. The results are presented in Table 4, below:

<table>
<thead>
<tr>
<th>University</th>
<th>N: Citations</th>
<th>Mean rank: Citations</th>
<th>N: Egghe’s g-index rat.</th>
<th>Mean rank: Egghe’s g-index (rat.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bucuresti</td>
<td>148</td>
<td>252.25</td>
<td>241</td>
<td>373.19</td>
</tr>
<tr>
<td>Cluj</td>
<td>83</td>
<td>332.33</td>
<td>129</td>
<td>515.25</td>
</tr>
<tr>
<td>Iasi</td>
<td>69</td>
<td>215.21</td>
<td>129</td>
<td>326.87</td>
</tr>
<tr>
<td>Timisoara</td>
<td>72</td>
<td>222.46</td>
<td>99</td>
<td>346.83</td>
</tr>
<tr>
<td>Craiova</td>
<td>38</td>
<td>167.57</td>
<td>59</td>
<td>357.25</td>
</tr>
<tr>
<td>Tg. Mures</td>
<td>45</td>
<td>35.31</td>
<td>63</td>
<td>88.48</td>
</tr>
<tr>
<td>Total</td>
<td>455</td>
<td></td>
<td>720</td>
<td></td>
</tr>
</tbody>
</table>

The results of both tests are significant at .000 level for both citations and g-index (a Chi-Square coefficient of 162.813***, df = 5), rational variant (a Chi-Square coefficient of 184.722***, df = 5), which demonstrates the robustness of the conclusion that there are indeed differences in scientific impact between the institutions with reference to the sum of ranks (Kruskal-Wallis test). We also computed a Median test for the ‘fair’ partitions, and the results are similar: there are differences also in terms of medians for both citations and the rational variant of the G-
index, and the tests are significant at the .000 level. Methodologically, the 'fair' partition is able to tap a significant share of the collective (institutions) level variance in scientific influence and thus to produce meaningful comparisons between institutions. Moreover, due to the similarity and quasi-linearity of the distributions we can control the within groups variance. Thus the 'fair' category could be considered representative for the collective level performance when aggregated from the level of individuals. It includes most of the research active academics whose performance are nonetheless not in the outstanding range and thus could be attributed to institutional rather than personal factors. Besides, it constitutes a robust measure that is independent of the size and shape of original distributions, being based on a parameter-free approach to citation-impact classification (Glänzel 2007, 2010).

Discussion
To resume our contribution, we used CSS, a parameter-free approach to citation-impact, to compare ‘inter-institutional’ performance in the condition of high ‘intra-institutional’ variance in six highly skewed distributions of bibliometric data associated with the medical schools in Romania. Our analysis represents a study case on a real population, which is primarily determined by tenure in the departments under research, while most of the populations of individuals studied in scientometric research are determined differently. Such populations, termed ‘bibliometric’, are inferred from the scientific outputs (articles) indexed in major databases (usually Thompson Reuters or Scopus datasets) based on the subject field taxonomies used by these databases (Albarrán et al. 2011; Albarrán, Perianes-Rodríguez, și Ruiz-Castillo 2015; Chatterjee, Ghosh, și Chakrabarti 2016; Perianes-Rodríguez și Ruiz-Castillo 2015). This strategy is common in the study of publication, as analysts "frequently lack[...j] information on the realities of labour, due to the absence of databases on the researchers, and on their institutional, discipline and field affiliations" (D’Angelo și Abramo 2015). However, the real population approach was used by other authors as well, see for example (Abramo, D’Angelo, și Di Costa 2008) when individual productivity was analysed according to the administratively determined area in which each surveyed researcher operated. The difference between the two types of populations is not only ontological, but has some important conceptual and methodological implications. The ‘real population’ we study is more heterogenous, as it includes academics with different profiles, obligations, attitudes and orientations. Thus, numerous academics are either not indexed in the reference database (Scopus), or uncited – almost half of the population, more precisely 47.87%, fall in this category. Usually these academics are not included in ‘bibliometric populations’. Furthermore, most of the cited academics have relatively low scores: only 2.34% of them have a value of Hirsch’s index which is greater than 5. Thus, the lower, most part of the distribution violates the condition of non-triviality (Bouyssou și Marchant 2014) because for more than half of the individuals, a ranking cannot be produced. However, we believe that such a characteristic of the population may not be exceptional, as similar difficulties are reported by other authors (Anania și Caruso 2013). Such an observation is backed up by empirical studies as well. For example Ruiz-Castillo & Costas (2014) found that 68% of authors in all fields published only one Web of Science indexed article over a period of nine years; their research spans over the period 2003-2011 and includes an impressive sample of 17.2 million disambiguated authors. The uncited academics may not be relevant for measures meant to quantify the ‘institutional quality’ through presence in an elite league of achievers, but their behaviour can be illustrative for their interaction with the institutional context. For example, not all of the uncited academics in the population we studied are new entrants i.e. generally assistants – see the left side of Figure 4. According to both accumulation effects and the normative definitions of tenure, the lower ranks academics are expected to be less prolific in terms of research results – this relationship is illustrated in the right side of Figure 4. However, we consider the high percentages of uncited
academics from the higher ranks indicate that impact is insufficiently incentivized. We presume that the proportions of uncited academics in departments are not accidental, nor random, but rather a reflection of the interaction between what Olson terms ‘internal organization’, ‘informal coordination’ and selective benefits, institutionalized through the policy context.

**Figure 4:** University ranks of uncited academics (left) and percentage of uncited academics within total, per university ranks (right)

Individual impact is evaluated only in relation to tenure, which is mostly determined through centralized standards, agreed by academic nominees. A brief inspection of the institutionalization of the professorship reveals that impact (citations, or citation derived measure) has not been traditionally included among the requirements for professors and associated professors – where national standards are in place since 2005. Furthermore, the tenure requirements which institute individual evaluations are only associated with promotion, not with day to day activity. This is an explanation of the fact that a significant share of professors and associated professors are uncited. Scientific impact is usually evaluated on a yearly basis only at the collective level, mostly for the institutional funding policy, but the aggregation rules are prone to ‘the exploitation of the great by the small’, or ‘free-riding’ phenomenon.

**Conclusion**

Kwiek (2015) argues that the achievements of the top 10% performers in a representative sample of academics in eleven countries in Europe (Romania not included) can be explained through ‘individual’, not ‘institutional’ factors. Though our research employs a different perspective and different metrics of performance, we consider that a similar interpretation can be proposed for the distribution. The achievements of the top academics, mainly concentrated from the 95th percentile to the high end of the distribution, can be seen as determined by agency, rather than by structure, whereas the achievement of the ‘fair’ category can be representative to ‘institutional’, structural factors.

We started our contribution from a particular problem – that of the obscured contribution of most of the academics within real populations to some influential collective research evaluation exercises. For this purpose we used conceptual tools which are specific for the theory of collective action and we analysed a real population of academics, within six medical schools – a rather limited sample, but exhaustively aggregated from the real population of Romanian medical schools.

However, in dealing with the particular situation of Romania, we touched upon some wider conceptual and methodological aspects related to the evaluation of research for policy purpose. We discussed in details methodological aspects regarding the quantification of quality in highly skewed populations. Without entering the lengthy debate regarding the universality of distributions (Chatterjee, Ghosh, și Chakrabarti 2016; Kaur, Radicchi, și Menczer 2013; Peterson, Pressé, și Dill 2010; Radicchi, Fortunato, și Castellano 2008; Waltman, Jan van Eck, și Van Raan 2012), the populations we studied have reasonably similar shape with those studied.
by other scholars. We also extended the initial scope of the CSS, on one hand, we used it on a rational variant of Egghe’s g-index, while it was traditionally employed for citations or publications. Our option was motivated by the similarity of the shapes of the distributions. On the other hand, we used it for a different purpose, that of identifying a measure of ‘inter-institutional’ variance among a set, be it limited to six distributions with very high ‘intra-institutional’ variance. In doing so, we touched upon an important methodological problem in the social sciences: the micro-macro transition, or how do we aggregate, in a meaningful way, properties of the individuals in order to make valid inferences at collective (social system) level. Our main findings relate to the properties of the ‘fair’ category within the CSS. These segments of the distributions include researchers whose scientific impact can be measured above a trivial level and, for our population, they exhibit two important features: linearity and proportionality. These properties make the ‘fair’ segments of the distributions amenable to comparing medians between institutions. In contrast, the "high" category distribution includes individuals that exhibit severe dispersion in terms of individual scientific impact. Such a behaviour of the populations reduces considerably the linearity of the data, and hence the relevance of aggregation using central tendency measures. Moreover, the aggregation of the 'high' categories are more sensitive to what economists term as ‘king effects’, which obscures the relation between the individual and the institutional level scientific impact.

Our main argument is that an aggregate level scientific impact, calculated as a population's central tendency measures on the basis of individual scores is methodologically more relevant, and also politically more appropriate when calculated on the basis of "fair" category. The study-case approach, though it was useful to shed light on what we consider an under-researched topic in scientometrics: that of the characteristics of real populations of academics, has obvious limitations in terms of the generalization of results. Further research should test the robustness of this measure by extending it to a larger population of universities. Although our study includes a high number of individuals, the corresponding collective entities are rather few. Extending the research to more universities and different subject fields will bring more important evidence on how aggregation from individual to collective level could be possible.

References
Crespo, Juan A., Yunrong Li, și Javier Ruiz-Castillo. 2013. „The measurement of the effect on citation inequality of differences in citation practices across scientific fields”. PLoS ONE 8(5).
The role of the h-index and the characteristic scores and scales in testing the tail properties of scientometric distributions. *Scientometrics* 83(3): 697-709.


Thelwall, Mike. 2016. „Are the discretised lognormal and hooked power law distributions plausible for citation data?” *Journal of Informetrics* 10(2): 454-70.


---

i Italics in the original text.


The granularity of disciplinary structures for benchmarking citation impact. The case of CSS profiles

Wolfgang Glänzel\textsuperscript{1}, Bart Thijs\textsuperscript{2}

\textsuperscript{1}wolfgang.glanzel@kuleuven.be

ECOOM and Dept MSI, KU Leuven, Louvain, Belgium

Dept Science Policy and Scientometrics, Library of the Hungarian Academy of Sciences, Budapest, Hungary

\textsuperscript{2}bart.thijs@kuleuven.be

ECOOM, KU Leuven, Louvain, Belgium

Abstract

In this paper we study the effect of granularity on Characteristic Scores and Scales (CSS). Unlike the traditional indicators that are mostly based on means and quantiles, CSS require the reduction of the citation distributions collaboration of the underlying reference population to four states (classes) and thus higher a different level of granularity. While the question of the choice of granularity is at higher levels of aggregation usually not critical since countries and university have rather multidisciplinary profiles, at lower aggregation levels specialisation becomes more typical. Inappropriate granularity might not warrant the depiction of the publication profiles at these levels in a correct and adequate manner and thus not add accurate citation profiles either.

In order to be able to process one complete annual volume of the Web of Science, we decided to calculate CSS thresholds and classes for two levels of granularity, namely sub-fields and WoS Subject Categories. With about 5\% deviation, we did not find a real significance. However, we identified journals with similar impact measures but different citation profiles, independently of the granularity.

Finally, we have pointed to the limitations in the choice of granularity – in terms of both too broad and too narrow subjects.

Conference Topic

Citation and co-citation analysis; The application of informetrics on evaluation; Indicators

Introduction

The question of granularity in the application of subject-based standards remains a constantly recurring issue. The gain of higher resolution and precision of details at zooming on the hierarchy of subject assignment is lost again by growing ambiguity and increasing fuzziness. Apart from randomness, Bookstein (1997) has identified two other demons being challenges to the measurement in informetrics and quantitative science studies. He pointed to fuzziness as the second demon reaching beyond probabilistic uncertainty as it is related to the impossibility to accurately represent concepts by traditional variable sets. Finally, ambiguity pushes the idea of fuzziness to its extreme: The conceptual basis for measurement itself is often weak although we aim at accurate measurement. These demons do also affect the possibility of accurate subject assignment in bibliometric applications as the conceptual basis underlying subject delineation and classification is often weak. In particular, we can identify two main weak points regarding subject classification. The first one concerns the way how subject assignment is made and the second one relates to the granularity of subject hierarchies. The need for accuracy is often confronted with the practical limitations as well. These limitations may results from indirect assignment (e.g., through journal classification), hard clustering or even by changes in context and perspective, that is, by the question of research in a special topic is still relevant for the subject in question or for the activity of a department under study. While at higher level of aggregation this effect may still be tolerable and benchmarking may still lead to acceptable results if high-level subject classification is used, for instance, in national comparison of publication activity and citation impact in fields like chemistry or biosciences, at lower levels of aggregation, at least at the level of scientific journals, university departments and research
teams this approach is stretched to its limits as the profiles of these units are very specific and require more individual approaches. Also different publication types might require different levels of granularity as Adams and Testa (2011) stressed in the context of the – at that time – newly introduced Book Citation Index (Thomson Reuters, now Clarivate Analytics – BKCI). At lower aggregation levels also first dramatic conflicts between subject delineation and classification become blatantly obvious. In other words, acceptable delineation is often not possible by using standard classification alone. Ideally quasi-unique paper-based assignment of high resolution would be a solution but in practice this might turn counterproductive as further zooming in may result in highly grainy blow-ups that may destroy the details, thus shapes and contours become fuzzy. The desire to correct ambiguity and fuzziness and to sharpen the blurred picture by supplementary but arbitrary human decisions and selections resembles a bit the allegory of reality, dream and interpretation shown using the example of the photographer’s interaction with material and technology in Michelangelo Antonioni’s film “Blow-Up” (1966), where the attempt to blow up a photograph to gain more information and lucidity and to get evidence finally resulted in the destruction of the detail structure just leaving even more room for obscurity, imagination and speculation (cf. Lehman, 2013).

Coming back to bibliometrics, we will study the effect in borderline applications. We have used two levels of granularity for achieving this objective, the level of scientific journals and the Web of Science Subject Categories.

**General methodological issues**

The question of granularity has been raised and analysed in recent studies. Among those studies we would like to bring out three articles that have been devoted to the analysis of the stability of relative indicators in the light of granularity. The authors of the first one (Zitt, Ramanana-Rahary & Bassecoulard, 2005) observed a certain “instability of impact measures should not be interpreted in terms of lack of robustness but rather as the coexistence of various perspectives each having their own form of legitimacy”. In default of ‘one best level’ of observation they therefore pled for particular prudent in interpretation of citation indicators instead of just choosing or preferring one of the possible worlds. Adams, Gurney & Jackson (2008) have tested the hypotheses stated by Zitt et al. using the example of the 2001 UK Research Assessment Exercise (RAE) and found that the correlation between average normalised citation impact and peer-reviewed grade varied according to the hierarchical selected level. Both studies use journals and different hierarchical levels of subject classification. When interpreting the results, we have to take two important issues into account. The first one is simply a methodological one; journals form a partition, while subject classification do not. Subjects always overlap to a certain extent, where the extent depends on the hierarchical level and, of course, on the underlying scheme chosen for the exercise as well. In order to overcome the usual challenge of overlapping subject categories and disciplines, Zitt et al., have forced single subject assignment of journals to subject categories and of subject categories to disciplines thus reaching partitioning the document space in all cases. Although this seems to be quite reasonable and convincing at the first sight, we have to point to the following crucial issue. The second issue is much more critical as it is of conceptual nature. Although in the Clarivate Analytics Web of Science Core collection (Wos) and the Scopus databases subject fields are defined on the basis of journal assignment, there is a fundamental difference between journal- and subject-based standards. Authors are bound up with their specialty due to education, their skills and career choices, but the discipline is not individually selected for each publication. Yet in the case of journals authors do decide where to publish their manuscript. Thus journals always reflect certain aspects of publication strategies since the authors have the – admittedly sometimes limited – freedom to choose the journal standard along with the scope for the target audience. In order to illustrate this, we refer to our plain old indicator triplet consisting of the mean observed, (journal) expected and field expected citation rates (MOCR, MECR and FECR, cf.
Glänzel et al., 2009). MECR and FECR provide two different expectations and the comparison of these two baselines reveal interesting details about the analysed unit’s publication profiles and strategies. For instance, MOCR > MECR > FECR means that the factual citation impact is above both expectations, where scientists affiliated with this unit, on an average, publish in journals with citation impact higher than the field standard, while MECR > MOCR > FECR means that the unit exhibits a factual citation impact above the expected subject-based standard and it does publish in high-impact journals but this standard is, on an average, not reached. What remains, is the comparison of granularity effects at different hierarchical level indeed. In 2009 we have made therefore an experiment using the example of 676 European universities with at least 50 publications each in the period 1999–2001 that have been selected from a total of roughly 2000 institutions (see Glänzel et al., 2009).

The three-year citation-window based field-expected citation rate (FECR) was determined on the basis of different hierarchical levels of subject classifications, the WoS subject categories, the 60 sub-fields and 15 major field according to the Leuven-Budapest subject classification...
scheme, which is built upon the WoS categories (cf. Glänzel and Schubert, 2003). Two of the results are shown in Figure 1. NMCR stands for the ratio of observed and field-expected citations rate using different granularity.

According to the expectations, the correlation between the ECOOM subfields and the WoS subject categories proved very strong \( r^2 = 0.971 \), while the correlation between the highest (major fields) and lowest hierarchical level (subject categories) was distinctly weaker, but still quite strong \( r^2 = 0.844 \). This goes with a larger variance and, most notably, the large number of outliers in the bottom diagram might cause problems in practice. The most significant observation is, however, the closeness of the slope to the value 1.0. This means that the choice of the hierarchical level, i.e., the granularity, has practically no “scale-effect” on the relative indicators. In the present study we repeat this exercise in the context of the Characteristic Scores and Scales (CSS), which can be considered a distributional extension of the notion of ‘relative indicators’. Also here we applied both the ECOOM subfield level and the original WoS subject categories. Before we present and discuss the result, we briefly introduce the rudiments of the CSS method.

The granularity of CSS – Methods and results

1. On the granularity of indicators at the meso level

As we have mentioned in the previous section, CSS can be considered a non-linear (distributional) extension of field-standardised citation rates. The basic idea is to replace citation indicators by citation profiles that are based on a given number of “performance” classes. The idea behind the method was two-fold, firstly, to favour a parameter-free solution for the assessment, that should be insensitive to outliers, subject-specific peculiarities, publication time and citation window, and provides unique assignment of publications to classes, and, secondly, to overcome the traditional linear projection of a multi-faceted, multi-dimensional reality. The “reduction” of the original citation distribution to performance classes as proposed by Glänzel & Schubert (1988) seemed to be the ideal way to achieve this objective. Since then a number of studies has analysed the properties of CSS and confirmed the robustness of the method. An analysis of the granularity has, however, not yet been conducted. However, as CSS has recently been applied also to lower levels of aggregation, notably to institutions, journals and individuals, this might be critical because journals, research teams, and individual scientists might have specific publication profiles that are not sufficiently or adequately covered at the level of standard disciplines. Before we analyse the effect of the choice of the hierarchical level of subject assignment for the calculation of CSS classes, we briefly recall their definition and basic properties.

In a nutshell, characteristic scores are obtained by iteratively calculating the mean value of a citation distribution and subsequently truncating this distribution by removing all papers with less citations than the conditional mean. The number of profile classes used for the evaluation is the only “arbitrary” number that needs to be defined in the beginning. Usually four classes are sufficient. In order to obtain fours classes, the process is stopped after three iterations. The four classes stand for ‘poorly cited’ (Class 1), ‘fairly cited’ (Class 2), ‘remarkably cited’ (Class 3) and ‘outstandingly cited’ (Class 4) papers. Papers in class 3 and 4 can be considered highly cited. Although the method is based on mean values, the huge number of publications underlying the base-line guarantees that the influence of outliers remains marginal. A further issue in comparative assessment at practically any level of aggregation arises from the peculiarities of scholars’ communication behaviour in their research domains and disciplines. This results in different disciplinary standards of publication activity and citation impact. In order to be able to apply the method to the assessment of national research output and to multidisciplinary universities, one has to apply kind of “fractional” normalisation, when calculating the threshold values for class assignment. The detailed description of the procedure
and its mathematical background is documented in previous studies by the authors (e.g., Glänzel, Thijs & Debackere, 2014).

Just as all citation indicators, the CSS scores, too, are strongly dependent of the subject matter, but the distribution of papers over the citation classes defined as the intervals between these scores proved largely independent of the subject, publication year and citation window (Glänzel, 2007; Glänzel, Thijs & Debackere, 2014). Although CSS classes are not directly linked to percentiles, notwithstanding that, the baseline derived from the complete population provides a distribution of papers over classes of about 70% (Class 1), 21% (Class 2), 6%–7% (Class 3) and 2%–3% (Class 4), independently of publication year and citation window. Also Albarrán & Ruiz-Castillo (2010) found a similar 70–21–9 rule when they combined the two upper performance classes for papers published between 1998 and 2002 and using a 5-year citation window. This robustness has made the method attractive as universal bibliometric profiling tool.

Before we further elaborate the CSS scheme for studying the effect of granularity, we go briefly back to a more general question. The research profiles of the units of macro- and meso-level analysis reflect multi-disciplinary or at least activities in broader subjects. The effect of specialisation, even within subject areas that are, otherwise, relevant for the units under study, becomes more pronounced at lower levels of aggregation and most notably at the micro level. The specialisation of individual scientist as reflected by their publication profiles can often hardly be captured by pre-defined disciplines or subject categories. The same applies to research topics and sub-disciplines. Scientometrics as a topic or discipline might serve just as an example: The discipline is, on one hand, a highly specialised sub-discipline within information & library science, and, being a truly inter-disciplinary field, is, on the other hand, related to many other fields as well and its metrics thus need to be compared with various different subject standards. The relationship with other disciplines is established cognitively by the actual application and practically through the journal assignment to the relevant subject categories. However, none of most subjects outside information science is representatively covered by scientometrics. The topic standard may also be distinctly different from that of its mother discipline meaning that gauging literature in a sub-discipline against the main-discipline standard would result in a distorted picture. In the case of scientometrics, actually a higher standard has to be assumed in order to obtain a realistic benchmark (see Glänzel, Thijs & Debackere, 0000). These are the reasons why the application of subject classification schemes might fail if applied as a basis for the benchmarking specialised topics and the citation impact of individual scientists. The same effect of incomplete or not representative coverage may occur at the meso and micro level, if (specialised) subsets of disciplines do not necessarily represent the standard of the discipline. In both cases the question of the granularity of the underlying classification system emerges. Even, if ambiguity is sufficiently compensated by subject fractionation and weighting (see Glänzel, Thijs & Debackere, 2014), unlike in the case of higher aggregation levels such as universities and countries (cf. Figure 1), the granularity might not prove appropriate to provide results of acceptable accuracy if we break down data to the topic, journal or team level. This is the reason why we decided to use an experimental approach to adjusting the “CSS resolution” for subject/topic selection in journal analysis. In order to be able to conduct large-scale analyses, we will study the effect of granularity at the journal level. This allows us to process all data of a complete volume of the WoS data base including the SCIE, SSCI and A&HCI editions in a systematic way.

For the present study we have chosen the complete 2013 volume of the WoS journal editions. We have only taken so-called citable items, that is, documents of the types: article, letter and review, into account. Again we have used the two hierarchical levels, the 74 ECOOM disciplines and the roughly 250 WoS subject categories we have used for subject assignment. First we calculated field standardised citation rate for journals according at both levels of
aggregation. Quasi as a by-product, we obtain a subject-insensitive citation metric that would allow journal ranking across subject fields. Yet, our intention is to overcome linearity in depiction citation impact, therefore we use these measure just as auxiliary tools. Figure 2 shows the plot of subject-normalised mean citation rates for the three-year citation window 2013–2015 based on ECOOM subfields vs. WoS Subject Categories. Only journals with at least 10 papers in 2013 are plotted and analysed. Since many journals are rather small, i.e., more than half the journals published less than 100 citable papers in 2013, we expected a larger variance and outliers being more frequent than in the case of universities. However, the correlation still proved to be strong and the slope of the linear regression is again close to the value one, substantiating that the level of subject assignment does not affect the scale of the standardised citation index.

![Plot of subject-normalised mean citation rates based on ECOOM subfields vs. WoS Subject Categories for 11964 journals covered by the WoS database](Data sourced from Clarivate Analytics Web of Science Core Collection)

In the second step we calculated the CSS class distributions for the complete population, that is, for all citable papers indexed and all journals covered in the 2013 WoS volume. The ECCOM classification resulted in the class distribution of 70.0% (Class 1), 21.4% (Class 2), 6.2% (Class 3) and 2.4% (Class 4), while the corresponding distribution based on WoS categories 69.8%, 21.4%, 6.3% and 2.5%. No statistical test is needed to understand that the deviation of the two distributions from each other is negligible. Nevertheless, we applied a $\chi^2$ homogeneity test to test the deviation of class distributions to all individual journals with at least 10 citable papers in 2013. Using the critical value of 7.81 at a confidence level of 0.05, we found that 94.6% of all journals do not differ significantly if the two levels of subject assignments are applied, while deviation of 5.4% of the journals must be considered significant. This is practically in line with the assumed confidence level.

2. Characteristic Scores and Scales in journal assessment

From the historical viewpoint, the use of the CSS in the context of journal assessment was the first application and this was actually the application the method was designed for (see Glänzel & Schubert, 1988). Unlike subject assignment or the profiles of individuals and research units or countries, journals do not require multiple assignments of papers. Journals form a true partition of the document space that is covered by a bibliographic database and also of any paper set under study. If papers are to be assigned to journals on the basis where they have been
published, assignment is always unique and thus no fractionation is needed. This property essentially simplifies the application of bibliometrics to journal indicators. On the other hand, journal analysis is still an important methodological topic in scientometric research and an indispensable fundament for supplementing bibliographic databases by journal metrics (cf. Journal Citation Reports (JCR), Scimago Journal Ranking (SJR), Scopus CiteScore metrics) and as long as journals serve as the basis of subject classification, journal indicators remain essential issues in bibliometric studies.

Table 1. Examples of selected journal pairs with almost identical mean citation rates (MCR) but different CSS profiles in 2013 with three-year citation window, based on ECOOM subfields (1st row) and WoS Categories (2nd row) [Data sourced from Clarivate Analytics Web of Science Core Collection]

<table>
<thead>
<tr>
<th>Journal title</th>
<th>Statistics</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced Energy Materials (n = 157)</td>
<td>MCR = 31.92</td>
<td>4.5%</td>
<td>29.3%</td>
<td>33.1%</td>
<td>33.1%</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$ = 1.59</td>
<td>7.6%</td>
<td>29.9%</td>
<td>32.5%</td>
<td>29.9%</td>
</tr>
<tr>
<td>Molecular Biology and Evolution (n = 238)</td>
<td>MCR = 31.71</td>
<td>42.0%</td>
<td>42.0%</td>
<td>13.0%</td>
<td>2.9%</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$ = 0.00</td>
<td>42.0%</td>
<td>42.0%</td>
<td>13.0%</td>
<td>2.9%</td>
</tr>
<tr>
<td>AJRCCM (n = 460)</td>
<td>MCR = 15.84</td>
<td>42.4%</td>
<td>20.4%</td>
<td>23.0%</td>
<td>14.1%</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$ = 3.94</td>
<td>44.8%</td>
<td>23.3%</td>
<td>21.5%</td>
<td>10.4%</td>
</tr>
<tr>
<td>BMC Medicine (n = 267)</td>
<td>MCR = 15.75</td>
<td>28.1%</td>
<td>52.8%</td>
<td>17.6%</td>
<td>1.5%</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$ = 7.66</td>
<td>28.1%</td>
<td>61.0%</td>
<td>10.1%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Scientific Reports (n = 2455)</td>
<td>MCR = 11.16</td>
<td>61.1%</td>
<td>31.7%</td>
<td>5.8%</td>
<td>1.4%</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$ = 0.00</td>
<td>61.1%</td>
<td>31.7%</td>
<td>5.8%</td>
<td>1.4%</td>
</tr>
<tr>
<td>Nutrition Reviews (n = 83)</td>
<td>MCR = 11.16</td>
<td>21.7%</td>
<td>38.6%</td>
<td>27.7%</td>
<td>12.0%</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$ = 6.15</td>
<td>38.6%</td>
<td>26.5%</td>
<td>22.9%</td>
<td>12.0%</td>
</tr>
<tr>
<td>Journal of Lightwave Technology (n = 578)</td>
<td>MCR = 5.60</td>
<td>51.6%</td>
<td>29.4%</td>
<td>13.8%</td>
<td>5.2%</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$ = 2.53</td>
<td>51.6%</td>
<td>32.5%</td>
<td>11.4%</td>
<td>4.5%</td>
</tr>
<tr>
<td>Clinica Chimica Acta (n = 364)</td>
<td>MCR = 5.60</td>
<td>72.0%</td>
<td>22.3%</td>
<td>4.4%</td>
<td>1.4%</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$ = 16.08</td>
<td>58.8%</td>
<td>30.8%</td>
<td>6.3%</td>
<td>4.1%</td>
</tr>
<tr>
<td>International Journal of Number Theory (n = 120)</td>
<td>MCR = 0.90</td>
<td>74.2%</td>
<td>20.8%</td>
<td>4.2%</td>
<td>0.8%</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$ = 0.00</td>
<td>74.2%</td>
<td>20.8%</td>
<td>4.2%</td>
<td>0.8%</td>
</tr>
<tr>
<td>Advances in Materials Science and Engineering (n = 248)</td>
<td>MCR = 0.90</td>
<td>97.2%</td>
<td>2.8%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$ = 1.63</td>
<td>98.8%</td>
<td>1.2%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

In the first applications, CSS were applied to chemistry journals to gauge journals Mean Citation Rates against the CSS standards calculated for each subfield of chemistry. At that time the share of uncited paper was still used as a supplementary Class 0 (below the poorly cited articles). Later on we restrained from this solution because of lacking robustness: The share of uncited papers strongly depends on both the citation windows and the subject matter. The reason for the instability is that this is the only class that is defined on a fixed criterion (i.e., uncitedness), while all other classes are based on variable thresholds that depend on time and research topic. The basic idea at that time was to provide additional information about the journals’ position in a discipline beyond the usual ranking exercises. Schubert et al. (1989) already applied CSS to journals, now on the basis of two scores only and with the purpose to gauge national contribution to journal impact but not to compare class distributions by journals. Now we will use CSS class distributions as an alternative to journal ranking. This also demonstrates a new and completely different application context. Although this kind of application does not allow any linear ranking and the interpretation of CCS class distributions is not always straightforward, the added value of information and the more detailed picture of citation impact that we obtain compensates for the lack of simplicity. The surplus of information
results from two sources that have been addressed many times in the scientometric literature, first the subject-specific peculiarities of citation impact and, second, the various shapes of citation distributions underlying journal impact. Glänzel and Moed (2002) and Glänzel (2009) have given examples of journals with almost identical Mean Citation Rates but different distribution shapes. We have computed Mean Citation Rates (MCR) and CSS class distribution for all journals in 2013 using a three-year citation window. We have applied the fractionation process for subjects according to the procedure described in section 3.2 since many journals have multiple subject assignments. Table 1 gives the MCR values of 5 selected journal pairs with almost identical citation impact (MCR) each but distinctly different profile types according to their CSS class distributions. We have chosen journals representing different “standards” to illustrate that this phenomenon might occur in all impact classes ranging from high to low standard.

3. Granularity at the limits – journal-based benchmarking

In the previous two sub-sections we have applied bibliometric field-based indicators at two different levels of granularities to the profiling of journal impact. Experience showed, on the one hand, that broad fields, such as the 15 major fields according to the Leuven-Budapest classification scheme or the 22 science areas of Clarivate Analytics Essential Science Indicators (ESI) are less suited to serve as the basis of reference standards. In particular, the major fields and broad science domains combine theoretical, experimental, methodological, empirical and applied research in one field. In contrast to this high level of aggregation, Table 2 shows the aggregate impact factors of ten selected subject categories in physics in the 2015 science editions of the Journal Citation Reports. It is clear that there can hardly be found a common reference standard for, e.g., condensed matters (CMPh) and mathematical physics. The barrier for applying broad fields is therefore two-fold, we are faced with cognitive reasons (comparability – why compare mathematical physics or acoustics with CMPh?) as well as different metric standards (commensurability) if we choose to broad subject fields for subject normalisation.

Table 2. Aggregate Impact Factors (AIF) of ten subject categories in Physics according to the 2015 Sciences Edition of the JCR [Data sourced from Clarivate Analytics Journal Citation Report]

<table>
<thead>
<tr>
<th>Subject Category</th>
<th>AIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHYSICS, CONDENSED MATTER</td>
<td>4.023</td>
</tr>
<tr>
<td>PHYSICS, PARTICLES &amp; FIELDS</td>
<td>3.880</td>
</tr>
<tr>
<td>PHYSICS, APPLIED</td>
<td>3.261</td>
</tr>
<tr>
<td>PHYSICS, ATOMIC, MOLECULAR &amp; CHEMICAL</td>
<td>3.028</td>
</tr>
<tr>
<td>PHYSICS, NUCLEAR</td>
<td>2.788</td>
</tr>
<tr>
<td>OPTICS</td>
<td>2.220</td>
</tr>
<tr>
<td>CRYSTALLOGRAPHY</td>
<td>2.219</td>
</tr>
<tr>
<td>PHYSICS, FLUIDS &amp; PLASMAS</td>
<td>2.094</td>
</tr>
<tr>
<td>ACOUSTICS</td>
<td>1.853</td>
</tr>
<tr>
<td>PHYSICS, MATHEMATICAL</td>
<td>1.802</td>
</tr>
</tbody>
</table>

On the other hand, if the narrower fields or disciplines, such as the WoS subject categories are chosen, the already discussed issues of fuzziness and ambiguity challenge both methodology and its implementation. Multiple journals assignment to subjects becomes more frequent and results in sometimes exuberant fractionation and weighting that is needed to calculate and apply reference standards and CSS scores. Without this the application of such procedure the choice of an appropriate reference value would be impossible as has been shown using the example of
the journal *Bioinformatics* that took deviating positions in journal ranking according to the different WoS categories it was assigned to (cf. Glänzel, 2011). An arbitrary restriction to single assignment at this level would, however, hardly be justified from the cognitive viewpoint nor provide accurate reference standards for any indicators to be built on the assignment.

Recalling the results by Zitt, Ramanana-Rahary & Bassecoulard (2005), who referred to “the coexistence of various perspectives each having their own form of legitimacy”, we see a further limitation to increasing the resolution, even if we respect the legitimacy of different granularities. As we already have confirmed this notion in the context of relative mean citation rates in our section on general methodological issues, subject- and journal-based reference standards do reflect different perspectives. We have calculated the journal based relative citation rate (RCR) as well its sub-field based pendent (NMCR). The different concepts underlying these structures becomes apparent from the comparison of the national indicators that are given in Table 3. For this table we have compiled a list of the 25 most active countries.

**Table 3. Discipline (NMCR) and journal based (RCR) relative citation rates and CSS classes based on a 3-year citation window for the 25 most active countries in 2013 [Data sourced from Clarivate Analytics Journal Citation Report]**

<table>
<thead>
<tr>
<th>Country</th>
<th>Papers</th>
<th>Discipline based</th>
<th>Journal based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NMCR C1 C2 C3 C4 RCR</td>
<td>C1 C2 C3 C4</td>
</tr>
<tr>
<td>Australia</td>
<td>54,205</td>
<td>1.39 60.6% 26.4% 8.9% 4.2% 1.18</td>
<td>59.6% 25.3% 9.1% 6.0%</td>
</tr>
<tr>
<td>Austria</td>
<td>14,094</td>
<td>1.43 59.9% 26.7% 9.2% 4.3% 1.20</td>
<td>59.2% 25.1% 9.2% 6.6%</td>
</tr>
<tr>
<td>Belgium</td>
<td>21,039</td>
<td>1.55 58.3% 27.1% 9.7% 4.9% 1.25</td>
<td>58.8% 25.2% 9.3% 6.7%</td>
</tr>
<tr>
<td>Brazil</td>
<td>39,499</td>
<td>0.76 78.8% 16.1% 3.6% 1.4% 0.98</td>
<td>66.9% 22.1% 6.7% 4.2%</td>
</tr>
<tr>
<td>Canada</td>
<td>64,137</td>
<td>1.37 62.2% 25.4% 8.4% 3.9% 1.14</td>
<td>62.3% 23.8% 8.3% 5.6%</td>
</tr>
<tr>
<td>China</td>
<td>218,125</td>
<td>1.01 69.6% 21.3% 6.4% 2.7% 1.11</td>
<td>62.7% 24.4% 8.4% 4.5%</td>
</tr>
<tr>
<td>Denmark</td>
<td>15,625</td>
<td>1.64 55.2% 28.8% 10.6% 5.4% 1.28</td>
<td>57.5% 26.5% 9.5% 6.5%</td>
</tr>
<tr>
<td>France</td>
<td>72,824</td>
<td>1.33 63.1% 25.1% 8.0% 3.8% 1.11</td>
<td>63.0% 23.4% 7.9% 5.7%</td>
</tr>
<tr>
<td>Germany</td>
<td>102,566</td>
<td>1.37 61.2% 26.2% 8.6% 4.0% 1.14</td>
<td>60.8% 24.6% 8.8% 5.8%</td>
</tr>
<tr>
<td>India</td>
<td>54,083</td>
<td>0.77 76.2% 18.0% 4.4% 1.5% 1.03</td>
<td>65.4% 22.7% 7.7% 4.2%</td>
</tr>
<tr>
<td>Iran</td>
<td>26,665</td>
<td>0.74 77.1% 16.7% 4.5% 1.7% 1.16</td>
<td>62.3% 23.6% 8.6% 5.5%</td>
</tr>
<tr>
<td>Israel</td>
<td>13,169</td>
<td>1.36 65.6% 23.8% 7.2% 3.4% 1.08</td>
<td>66.7% 21.6% 7.0% 4.6%</td>
</tr>
<tr>
<td>Italy</td>
<td>64,334</td>
<td>1.26 62.4% 26.0% 8.3% 3.4% 1.16</td>
<td>59.9% 25.2% 8.9% 5.9%</td>
</tr>
<tr>
<td>Japan</td>
<td>79,048</td>
<td>0.95 73.6% 19.5% 4.9% 2.0% 0.95</td>
<td>68.6% 21.3% 6.6% 3.5%</td>
</tr>
<tr>
<td>Netherlands</td>
<td>38,227</td>
<td>1.61 55.6% 28.7% 10.5% 5.2% 1.22</td>
<td>58.8% 25.6% 9.3% 6.3%</td>
</tr>
<tr>
<td>Poland</td>
<td>24,081</td>
<td>0.83 76.2% 17.7% 4.2% 1.9% 1.08</td>
<td>63.6% 23.2% 8.1% 5.1%</td>
</tr>
<tr>
<td>Russia</td>
<td>29,447</td>
<td>0.61 84.4% 11.6% 2.8% 1.2% 1.05</td>
<td>66.8% 21.4% 7.4% 4.4%</td>
</tr>
<tr>
<td>Korea</td>
<td>50,720</td>
<td>0.95 72.8% 19.8% 5.4% 2.0% 0.99</td>
<td>66.3% 22.7% 7.3% 3.7%</td>
</tr>
<tr>
<td>Spain</td>
<td>56,839</td>
<td>1.25 64.8% 24.4% 7.5% 3.3% 1.11</td>
<td>63.1% 23.4% 7.9% 5.6%</td>
</tr>
<tr>
<td>Sweden</td>
<td>24,719</td>
<td>1.47 59.5% 27.2% 9.1% 4.2% 1.20</td>
<td>60.3% 25.0% 9.0% 5.7%</td>
</tr>
<tr>
<td>Switzerland</td>
<td>27,604</td>
<td>1.70 54.4% 28.6% 11.4% 5.6% 1.25</td>
<td>57.8% 25.6% 9.9% 6.7%</td>
</tr>
<tr>
<td>Taiwan</td>
<td>27,709</td>
<td>0.93 73.0% 20.0% 5.2% 1.9% 0.93</td>
<td>69.0% 21.5% 6.3% 3.2%</td>
</tr>
<tr>
<td>Turkey</td>
<td>27,276</td>
<td>0.64 81.7% 13.8% 3.3% 1.2% 1.04</td>
<td>66.4% 21.9% 7.1% 4.6%</td>
</tr>
<tr>
<td>UK</td>
<td>112,662</td>
<td>1.45 60.1% 26.4% 9.1% 4.4% 1.14</td>
<td>61.6% 24.1% 8.5% 5.8%</td>
</tr>
<tr>
<td>USA</td>
<td>386,556</td>
<td>1.38 60.7% 26.2% 9.0% 4.1% 1.07</td>
<td>62.6% 23.7% 8.3% 5.4%</td>
</tr>
<tr>
<td>Total</td>
<td>1,464,089</td>
<td>1.00 70.0% 21.4% 6.2% 2.4% 1.00</td>
<td>65.0% 22.6% 7.5% 4.9%</td>
</tr>
</tbody>
</table>

All citation indicators have been calculated on the basis of the 3-year citation window 2013–2015. The low NMCR values of India, Iran, Turkey, Russia and Brazil are contrasted by their
moderate or even high journal-based counterparts (RCR) indicating that these countries are publishing in rather low-impact journals. The direct comparison of the two underlying expected citation rates make this even more visible: For Iran, Turkey and Russia, the ratio of the journal and discipline based expectation ranges around 0.6. The comparison of the corresponding CSS classes confirms and deepens these observations. The first observation here concerns the baseline itself. The distribution of papers over CSS classes in the complete population changes towards a lower share of poorly cited and a consequently higher share of fairly and highly cited papers. This has also a strong effect mainly on countries with low impact: In countries with low citation impact, the share of poorly cited papers calculated from journal model is usually greater than the corresponding field-based share. Countries with generally high citation impact are not affected by this trend, on the contrary, as can been seen for Denmark, Switzerland, the Netherlands or Belgium. For these countries we also observe somewhat higher field-based than journal-based expectations. The ratio here amounts to about 1.3. In both extreme cases, we have looked at, we see the effect of national “publication strategies”, namely, in favour of publishing in low or high impact journals with respect to their corresponding subject standards. In this context we would like to recall the effect of choosing publication channels and fora already mentioned in the introduction.

However, apart from these perspectives a real barrier emerges, namely that given by the rules of mathematical statistics – at least in the context of CSS. And in this context we have to stress that the limitation applies to the population, not to the size of the sample units under study. The commonly accepted minimum for the application of mathematical statistics to mean-value and quantile based statistics lies around 30–40 items (cf. Vincze, 1974). WoS subject categories meet this criterion but about one quarter of all journals have less than 30 citable paper per year. For the calculation of reference standards for mean values and shares this is still acceptable. Now, analogously, the use of journals instead of subject categories or disciplines could be an interesting extension to journal-based expectation also in the framework of CSS. In other words, journal-based characteristic scores and scales could supplement the field based CSS giving another perspective to citation-class profiling. However, here the distribution based approach is stretched to its limits. For instance, if we assume a size of 40 items in the reference journal, the share of Class 4 items of 2.5% of the reference distribution would correspond to exactly 1 item. This might, however, challenge the robustness of the method. For this reason, we do recommend a selective extension of CSS to journal-based profile classes.

Conclusions

We have calculated characteristic scores and scales for journals covered by the Web of Science Core Collection at two levels of granularity, 74 disciplines and the WoS Subject Categories. The results of the overall comparison did not show any substantial deviations that would tell against the robustness of the method. The comparison of CSS and mean citation rates have again substantiated that citation impact is not one-dimensional and should not be mapped by linear functions. Journals may have very similar mean citation rates (impact factors, or more generally, impact measures), while their CSS class distributions distinctly differs, independently of the underlying granularity. Our findings are partially presented on a limited selection of journals. In an extended later version of this study we intend to present data on more journals in a supplementary online material according to the corresponding journals publication policies.

Unlike traditional mean-value or quantile based bibliometric indicators CSS needs somewhat larger reference sets to obtain robust reference distributions. The reason is the higher resolution of the underlying distribution based reference classes. According to the experience made in this study, subject categories already meet this requirement, journals are less suited to build the reference classes for characteristic scores and scales.
Summarising, the proper choice of granularity has two limits, both the heterogeneity of too broad subject fields and an increased resolution towards too narrow topics could affect the validity of results. As so often, when turning theory to practice, the optimum is somewhere in between.

Acknowledgement

Figure 1 has been reproduced from Glänzel et al. (2009) with permission of the publisher. Table 1 presents data on subfield-based journal indicators shared with figures of Table 9 in Glänzel, Thijs & Debackere (0000).

References


Glänzel, W., Thijs, B. & Debackere, K. (0000), *Citation classes: a distribution-based approach to profiling citation impact for evaluative purposes*. In: Glänzel et al. (Eds), Springer Handbook of Science and Technology Indicators, Springer Heidelberg, in preparation.


What factors are associated with more authors? A case study of Danish Economics

Abstract
The number of authors is increasing in the social sciences, but co-authorship studies often explain these increases in the number of authors by referring to authorship studies in the natural or life sciences or by using anecdotal evidence. Thus, a deeper understanding of why the number of authors is increasing in most field is necessary, as well, as whether there are changes in the research or in the researchers’ behavior. This study will examine the changes in research in Danish economics, and investigate to what extent these changes can be explained by external factors in the research culture. The study finds that articles with empirical research, quantitative research and/or survey have a slightly higher average number of authors than theoretical, interview, data and qualitative research. However, the average number of authors increases for articles with all types of research and research approaches, and it seems like something more than external factors are affecting researchers’ collaboration behavior.

Conference Topic
Science communication, Science of Science

Introduction
Research collaboration has become the norm in most research fields, and research is increasingly being communicated in co-authored peer-review research articles. Studies have shown that the average number of authors is rising in most research fields (Henriksen, 2016; Lariviere, Gingras, Sugimoto, & Tsou, 2015). The main focus is often on high energy physics and biomedicine with extreme co-authorship tendencies, where the norm in some areas is hundreds of authors and it is no longer uncommon with thousands of authors in the byline of an article (Cronin, 2001). The tendencies in these research fields often overshadow that the same tendency is occurring in other research fields, though it is occurring at a smaller scale in the fields of social science. Furthermore, these increases are not the results of a need for big laboratories or expensive equipment as in the natural and life sciences (Birnholtz, 2006).

Several studies have documented a steady increase in co-authorship in economics (Hudson, 1996; Laband & Tollison, 2000; Sutter & Kocher, 2004). The studies suggest various explanations for the increases based on general anecdotal evidence. For example, that a rapid growth of knowledge, increasing complexity of research, and interdisciplinary research have required more collaboration among researchers which the increase in co-authorship reflects.

These suggestions are aligned with conclusions from authorship studies in the health and natural sciences, which have found that the increasing more technical nature of the disciplines and larger data material would require more manpower and attribute to more labor-divided research tasks (Colledge, Elger, & Shaw, 2013; Laband & Tollison, 2000). These arguments thus suggest that the increase in co-authorship can be attributed to an increase in empirical studies based on quantitative methods and the uses of more advance statistical, while theoretical and qualitative work would be less likely to be co-authored authored (Corley & Sabharwal, 2010; Fisher, Cobane, Vander Ven, & Cullen, 1998). Furthermore, the general growth of the professions creates more opportunities to find suitable co-authors, as well as the lower travel costs and innovations in information and communication technologies make it easier to collaborate for distant researchers.

All of the above suggestions focuses on changes in conducting research, technological advances, research environment and travel costs. However, the changes in attitude to the requirements of co-authorship could also be a factor. For example, researchers’ have become quicker to require their names being added as co-authors when assisting others, exchanging data, mentoring, and engaging in collaborations, and reciprocating others contributions with offering of co-authorship. This can be seen as a natural consequence of the “publish and perish”
mantra that has been dominating academia in the last decades, and makes it necessary for researchers to add their names to more publications and improve the productivity on their resume. This mantra is also dominant in the mindset of funding agencies and research institutions, which encourage researchers to engage in collaborations, since collaborative research is perceived as being of higher quality, more visible, and more cited in a time of research publication production overload (Thelwall & Sud, 2016). Furthermore, research collaborations and co-authoring lower the risks of having a no-publishing period caused by negative results, data sampling problems or long peer review process.

This study addresses the suggestions by other studies to investigate deeper why there has been an increase in the number of co-authors per article and the share of co-authored articles. Previous studies often refer to how changes in research conduct and methods have affected the tendency to collaborate, which increase in co-authorship reflects. This tendency can be refer to as external changes in research culture, which are explicit and possible to observe, measure and document.

On the other hand, the internal changes in research culture affect researchers’ perceptions and practices of co-authorship, research collaboration, reward system, science communication system, and research responsibility. This study examines which kind of social science research articles have more co-authors, and if the increase in co-authorship corresponds to increase in certain research approaches and methods? These findings can contribute to a discussion of how external and internal changes in research culture can effect co-authorship tendencies.

Therefore, this study is a necessary step in the authorship studies to gain further knowledge about how the practice and frequency of co-authorship is changing. This study will address this lack of knowledge by examining the characteristics of economic articles, and thereby gain further knowledge about co-authorship practices. It begins with a review of how certain characteristics of the articles are related to co-authorship, followed by a methods and data section presenting the coding algorithm and the statistics. The results section presents first descriptive statistics followed by regression models. It concludes with a discussion of these results in relation to external and internal changes in research culture and the effect on co-authorship.

**Literature review of factors**

*Is theory the lone scholars’ game?*

Previous studies often find that empirical articles compared to theoretical articles are more co-authored (O'Neill, 1998; Sacco & Milana, 1984). Piette and Ross (1992) investigated 15 economics journals, and found that empirical work has a greater tendency to be co-authored. This corresponds to the suggestions that empirical research tasks sometimes requires researchers with different expertise and skills to collaborate, and the tasks can be divided among them (Barnett, Ault, & Kaserman, 1988). Furthermore, the development of online survey tools, easier access to data via online databases, and different software programs to analyze the collected material have made empirical work more accessible. Therefore, the study expects that an increase in co-authorship will have occurred over time in empirical research articles.

*Does the methodological approach affect the number of authors?*

The literature of social sciences co-authorship suggests that quantitative research compared to qualitative research are more likely to be co-authored, since it offers greater opportunities to work in a team-production model (Adams, Black, Clemmons, & Stephan, 2005; Ossenblok, Verleysen, & Engels, 2014). Few studies have further investigated the notion that quantitative research is more co-authored. Nowell and Grijalva (2011) used the Journal of Economic Literature (JEL) to determine if the publication is based on quantitative research, and find that quantitative research is not the only factor in the increasing likelihood of co-
authorship. Hunter and Leahey (2008) examined the key parts of the papers to assess if they contained quantitative or qualitative research, and they found that quantitative publications is more likely to be co-authored.

Quantitative research tasks are often easier to divide among team members, thus the tedious tasks of producing or/and cleaning data can be divided and executed by multiple team members. Furthermore, the data analysis often uses statistical methods, which can involve one or more team members. The data of qualitative studies often consist of interviews, focus groups, observations and/or text data, which may not always be easy to divide among members of a team, depending on the ontological and epistemological position. Therefore, this study expects that articles based on quantitative research methods instead of qualitative research methods will probably be co-authored.

Is the number of authors affected by whether the data is primary or secondary

The collection of data for a research project is often a labor-intensive task, and there are pros and cons depending on the kind of methodology and data collection usage in a research project. Particular quantitative research in economics uses secondary data, such as data from archives, databases or registers. Secondary data is probably easier to share among research collaborators, since the data is already available in archive etc. and this will encourage people to collaborate to obtain a synergy and more knowledge in the collaboration. It requires less work to collect, but may require a great deal of work for preparing such data for analysis. On the other hand, the labor-intensive tasks of collecting data can require and be divided among multiple people engaged in collaborative research project, which will be reflected in the number of co-authors of a publication.

These different aspects and advantages of primary and secondary data both give indication of how and why there can be more authors on a publication. It is therefore difficult to assess whether one can expect differences in how many authors of an article depending on the kind of data. The previous studies of authorship and research collaboration do not provide a clear answer to this question. Fisher et al. (1998) suggest that both large-scale primary data collection efforts and the increasing availability of large secondary datasets in archival depositories have increased the proportion of co-authored publications. However, they only find 3.1% of the political science articles in their sample using primary data, so they excluded the results from their study. Hunter and Leahey (2008) investigated whether the usage of primary data vs. secondary data is a factor in the increase of co-authored sociology articles. Their hypothesis is that secondary data are easier to share when collaborating, and they expect that it is more likely to be co-authored. However, they found that the proportions of co-authored articles have increased for both articles with primary and secondary data, and that in 2005 it is at the same level. This study examines the association of secondary data with a higher number of authors.

Do international and interinstitutional co-authored articles have a higher number of authors?

The increase in co-authored publications is found to contribute to an increase in international collaborations. The share of internationally co-authored publications doubled between 1990 and 2000 in the sciences, and at a more rapid speed than national co-authored publications (Nomaler, Frenken, & Heimeriks, 2013; Wagner & Leydesdorff, 2005). However, if one limits the study to solely the social sciences the findings are mixed. Lariviere, Gingras, and Archambault (2006) and Henriksen (2016) found that there has been an overall increase in international co-authorship in the social sciences, while Hunter and Leahey (2008) do not find any support for an increase in international co-authorship in sociology, they do find a slight increase in interinstitutional collaboration.

One of the reasons that international and interinstitutional co-authorship have increased could be due to the electronic development that has made it easier to interact and work with researchers from other institutions and countries. However, studies have shown that researchers
still have a tendency to collaborate with people in a close geographical proximity (Wagner & Leydesdorff, 2005). Perhaps because face-to-face meetings still are necessary, as well as neighboring countries often have more shared cultural values that make it easier to collaborate. This study focuses on Danish economic research, and since Denmark has a very small population, it would be normal for the researchers to find collaboration partners outside of their own institution and country. Three variables including the geographical distance (GCD), the international collaboration (IN), and the interinstitutional Danish collaboration (II) are used to investigate if the mean number of authors are higher because of collaborations outside one’s own institution.

*Normalized Journal Score NJS*

Previous studies of co-authorship and research collaboration often focus of the relationship between more co-authors and citations, and whether there is a connection between number of authors and the journal impact factor (JIF). There are very different view on whether there are a positive or negative relationship between number of authors and JIF. Studies propose that more co-authors shows that the research is of higher quality, thus researchers will strive to publish in journals with a high JIF. Hence, it will give the article a pre-quality seal of approval, more visibility (more citations), more prestige, and in some cases give the researchers a financial bonus.

Other studies, that found a negative correlation between the JIF and the number of authors, suggest that researchers will prefer to publish alone in high JIF journals to secure that the citations and credit do not have to be shared (Rutledge, Karim, & Reinstein, 2011). However, this argument assumes that researchers decide to be more inclusive in the selection of who fulfills the authorship requirements, regardless of the researchers’ contribution, as well as requiring a higher standard for who can be included as an author if they deem the article to be able to be published in a high JIF journal. This study is dealing with a social science fields, so the measurement of the journal will instead be the Normalized Journal Score NJS from CWTS Journal Indicators.
Methods and data:
All research articles with at least one Danish address belonging to Economics subject category were downloaded from the Web of science Social Science Citation Index on 19 October 2015. The period spans 35 years from 1980 to 2014. The sample contains 3,157 Economic articles. The data are divided into different time periods to show the development in co-authorship and changes in the article characteristics.

Data coding
Bibliographic information about the articles were downloaded in the original sample. This was followed by a retrieval of and matching data about authors’ institutional and national affiliation. The variables international (IN) and interinstitutional (II) collaboration was dummy coded. Followed by the calculation the variable GCD, NJS and interdisciplinarity. However, the coverage of social science references before 1995 are scarce, and the interdisciplinarity variable can only be used in the analysis of data after this year, and will only be included in the last regression model.

The publication content data was manually coded with regard to the content of the publications. This required accessing the full-text of each article. After the full-text of the article was located, a strict algorithm was used to assess the content of the article: First, it assessed whether the article were theoretical or empirical by examining the method or data section for usage of data or data collection. Second, if the article is empirical, it is further coded depending on whether it is quantitative, qualitative or mixed method. If it is assess to be quantitative if it uses survey, secondary data (register or archival data), and experiments. If the article uses interview, ethnographic or anthropological methods or documents in their study it is assess to be qualitative, and if it uses both quantitative and qualitative methods it is mixed method. Third, the usage of interviews, surveys, secondary data, document data or other are coded. All of the content variables are dummy coded, for example 1 for survey and 0 for not survey.

Regression count models
This study uses Poisson regression (Poisson) and negative binomial regression (NBR) count models to examine how the different factors influence the number of authors. The reason for using two different count models is because dispersion for the outcome count variable “number of co-authors per article” varies across fields and sub-periods, in some cases being over-dispersed and in others having a dispersion parameter almost equal 1. Thus, when it is over-dispersed it fits the criteria for a NBR model, and when the dispersion parameter is almost equal 1 or the variance and mean number of authors do not differ, it fits the criteria for a Poisson model. Because the analysis uses count models, the exponent of the different coefficients is calculated to make a more clear interpretation of the results. Thus, if all the variables are kept constant in the count model, then the exponent are the changes in the dependent variable (number of co-authors).

The formula for count models can be written as

$$N_{co-\text{authors}} = \text{Intercept} + \beta_1 x + \beta_2 x + \beta_3 x$$ ...

The previous studies of co-authorship emphasize multiple factors associated with more authors or co-authorship tendencies, thus this study will use different models to investigate what factors are associated with more co-authors, since some of the factors overlap. Each regression model will be presented for the overall period and for the three time periods (1980-1994, 1995-2004, and 2005-2014). The first period focuses on the pre-internet period, followed by a period with better long distance communication possibilities as well as cheaper travel opportunities. Last is a period, where in Danish context there have been a greater tendency to focus on research performance and new public management in the university sector. The variables NJS, GCD, IN and II are included in all of the models, and will therefore only be thoroughly discussed the first time they are presented in the models.
**Results**

The study finds that the productivity of Danish researchers in international journals increases during the last 35 years, as well as the total number of both theoretical and empirical articles increases, which corresponds with the general increase in the number of researchers and productivity in that period. The graphs in figure 1 and 2 show how co-authorship, the mean number of authors and the number of articles are distributed over the 35 year time period. The mean number of authors for the total number of articles increases by 98.41% (1.26 authors to 2.50 authors). In the first period, the mean number of authors and the median number of authors are similar for theoretical and empirical articles. However, this changes over time and the mean number of authors increases for empirical articles by 111.2% (1.25 authors to 2.64 authors) and for theoretical articles by 70.63% (1.26 authors to 2.13). However, the median number of authors continues to be similar for both kinds of articles.

*Figure 1: The evolution of the mean number of authors and the number of economic articles (N authors <= 10)*

The descriptive results for economics show that the share of articles with empirical research has consistently increased from 30.5% in the 1980s to 71.1% in 2010s (see figure 1 and 2), as well as the total number of articles have increased rapidly over the years. The development shows a shift in Danish economics research from being a theoretical focused research field to a more empirically focused research field. The rise of co-authored articles are similar in both theoretical and empirical economics.
The majority (90%) of empirical articles uses quantitative methods, while the last 10% is a mix of qualitative methods and mixed methods. Because of the great overlap between quantitative methods and empirical articles the descriptive statistics are fairly similar; thus they show that the share of co-authored articles goes from approximately 27% to 85%. Few of the economic articles uses qualitative research, but those that do, show that there are clearly differences in the share of co-authored articles as well as that the mean number of authors are lower for qualitative articles during the 35 years.

Figure 3 show the development of mean number of authors and number of articles depending on research methods. The mean number of authors have increased for all types of research, but mostly for articles using survey studies. The mean number of authors is very high for survey studies in the period 1995 to 1999, which is caused by a low number of articles and two articles being with 10+ authors. However, in the last period there are 144 articles with survey studies, where 91% of these are co-authored, so this to some extent confirms that using this kind of research method corresponds with working in larger team. Thus, the tasks involving
conducting a survey can be allocated to different team members, and thereby increasing the number of people involved. Interestingly, interviews studies have also increased in both number of articles and mean number of authors, perhaps because different tasks such as the designing, interviewing, transcribing, and analyzing can be assigned to different team members. Furthermore, the usage of secondary data is the kind of method that has increased the most in number of articles, but also by more than 1 author in the mean number of authors. This is perhaps because the usage of data from external sources makes it easier to share the data and the burden of data cleaning. Not to mention that some of these articles use and collect data using more than one method; the articles with 24 and 26 authors in the byline have both conducted a survey and used data from a database. This could require researchers with different skills and therefore increase the number of authors on the final product.

**Regression models**

The first model in Table 4 is based on the formula below:

\[ N_{co-authors} = \text{Intercept} + \beta_1 PG + \beta_2 NJS + \beta_3 GCD + \beta_4 IN + \beta_5 II + \beta_6 \text{Empirical} \]

<table>
<thead>
<tr>
<th>Table 1 Economics regression count model: Research approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent variable: No. of co-authors</td>
</tr>
<tr>
<td>All years (NBR)</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>NJS</td>
</tr>
<tr>
<td>GCD</td>
</tr>
<tr>
<td>IN</td>
</tr>
<tr>
<td>II</td>
</tr>
<tr>
<td>Empirical</td>
</tr>
<tr>
<td>No. of observations</td>
</tr>
<tr>
<td>1980-1994 (Poisson)</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>NJS</td>
</tr>
<tr>
<td>GCD</td>
</tr>
<tr>
<td>IN</td>
</tr>
<tr>
<td>II</td>
</tr>
<tr>
<td>Empirical</td>
</tr>
<tr>
<td>No. of observations</td>
</tr>
<tr>
<td>1995-2004 (NBR)</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>NJS</td>
</tr>
<tr>
<td>GCD</td>
</tr>
<tr>
<td>IN</td>
</tr>
<tr>
<td>II</td>
</tr>
<tr>
<td>Empirical</td>
</tr>
<tr>
<td>No. of observations</td>
</tr>
<tr>
<td>2005-2014 (Poisson)</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
</tbody>
</table>

1208
In the first period, the model indicates that NJS have small positive effect on the number of co-authors. If the article is empirical, it has 7% less co-authors compared to theoretical articles. The results correspond with the fact that 67% of the articles are theoretical in the first period, and the mean being similar for theoretical and empirical articles. The GCD do not seem to have any effect on the number of co-authors, while the variables IN and II are very influential, so an international co-authored article will have on average 8.70 times more co-authors, and an interinstitutional Danish article will on average have 7.09 times more co-authors than an article with only internal collaboration or no collaboration. The results for the variables NJS, GCD and empirical are not so robust when examining the confidence interval and p-value, while the IN and II seems reliable.

In the second period, IN and II have 4.79 and 3.30 times average more co-authors than non-international and interinstitutional Danish articles. The coefficient of the variable NJS becomes negative, while GCD seems unchanged. The empirical articles show 1.80 times more co-authors than theoretical, which fits the results from figure 2, where the mean number of authors for empirical articles is higher than theoretical. In the third period, NJS show 1.03 times more authors, meaning it has a very small influence on the number of co-authors. If the article is empirical, it is equal to 1.38 times more co-authors. IN and II articles show 3.46 and 2.94 times more co-authors. In the last two periods, the confidence interval and p-value for NJS shows that the results are not reliable, while the rest of the results seems robust. The results from the whole period show that international or interinstitutional Danish co-authored articles in general have a higher number of co-authors

The coefficients for the variables IN and II are very high, and especially IN are influence by being highly correlated (0.65) with co-authorship. Hence, the study examine whether these variables still would be influential if the sample consisted of solely co-authored articles. The model was recalculated, and even though the coefficients are lower, the variables remain influential one the number of co-authors, since IN now gives approximately 1.3 times more authors in the whole period and in each period, and II gives between 1.1-1.3 times more authors.

The first model shows that international and interinstitutional Danish collaborated articles have tendency to have more co-authors in the entire period. However, the GCD shows that these international co-authored articles probably are from neighboring countries. From the first period to the second in the model is there a clear shift in empirical articles being more influential on the number of co-authors, though the coefficient becomes smaller in the last period. This is align with the results for the whole period that confirms that empirical, international and interinstitutional Danish collaborated articles have a positive influence on the number of co-authors.

The model was recalculated with the variable quantitative instead of the empirical variable using the formula below:

\[ N_{co-authors} = \text{Intercept} + \beta_1 \text{MJNS} + \beta_2 \text{GCD} + \beta_3 \text{IN} + \beta_4 \text{II} + \beta_5 \text{Quantitative} \]

There are a 0.76 correlation between the empirical and quantitative variable for economics articles, and the results for this model are very similar to the first model. Hence, the quantitative articles have approximately 1.42 times more authors for the whole period. In comparison, the empirical articles have 1.49 times more co-authors than theoretical articles. Further details can be provided by the author.
The final model includes variables focusing on the research methods used in the articles:

\[
\log(N_{\text{authors}}) = \text{Intercept} + \beta_1PG + \beta_2NJS + \beta_3GCD + \beta_4IN + \beta_5II + \beta_6Empirical + \beta_7quantitative + \beta_8Data + \beta_9Interview + \beta_{10}Survey
\]

### Table 2: Economics regression count model: Research methods

<table>
<thead>
<tr>
<th>Dependent variable: No. of co-authors</th>
<th>$\beta$</th>
<th>$\exp(\beta)$</th>
<th>$z$</th>
<th>$P&gt;z$</th>
<th>95% Conf. Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>All years (NBR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.1909</td>
<td>0.3040</td>
<td>-24.4700</td>
<td>0.0000</td>
<td>-1.2863 - 1.0955</td>
</tr>
<tr>
<td>NJS</td>
<td>0.0295</td>
<td>1.0299</td>
<td>1.5400</td>
<td>0.1180</td>
<td>0.0000 0.0669</td>
</tr>
<tr>
<td>GCD</td>
<td>0.0000</td>
<td>1.0000</td>
<td>4.7700</td>
<td>0.0000</td>
<td>0.0000 0.0000</td>
</tr>
<tr>
<td>IN</td>
<td>1.5219</td>
<td>4.5809</td>
<td>29.0000</td>
<td>0.0000</td>
<td>1.4190 1.6248</td>
</tr>
<tr>
<td>II</td>
<td>1.2597</td>
<td>3.5244</td>
<td>18.3400</td>
<td>0.0000</td>
<td>1.1251 1.3944</td>
</tr>
<tr>
<td>Data</td>
<td>0.2883</td>
<td>1.3341</td>
<td>7.5500</td>
<td>0.0000</td>
<td>0.2135 0.3631</td>
</tr>
<tr>
<td>Interview</td>
<td>0.2546</td>
<td>1.2900</td>
<td>2.8200</td>
<td>0.0050</td>
<td>0.0776 0.4317</td>
</tr>
<tr>
<td>Survey</td>
<td>0.5718</td>
<td>1.7715</td>
<td>9.9600</td>
<td>0.0000</td>
<td>0.4593 0.6843</td>
</tr>
<tr>
<td>Document</td>
<td>-0.2237</td>
<td>0.7995</td>
<td>-1.7300</td>
<td>0.0800</td>
<td>-0.4770 0.0296</td>
</tr>
<tr>
<td>No. of observations</td>
<td>3,157</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1980-1994 (Poisson)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.7012</td>
<td>0.1825</td>
<td>-14.3000</td>
<td>0.0000</td>
<td>-1.9344 - 1.4679</td>
</tr>
<tr>
<td>NJS</td>
<td>0.0358</td>
<td>1.0364</td>
<td>0.5700</td>
<td>0.5670</td>
<td>0.0001 0.0001</td>
</tr>
<tr>
<td>GCD</td>
<td>0.0000</td>
<td>1.0000</td>
<td>-0.5900</td>
<td>0.5550</td>
<td>-0.0001 0.0001</td>
</tr>
<tr>
<td>IN</td>
<td>2.1723</td>
<td>8.7786</td>
<td>11.1300</td>
<td>0.0000</td>
<td>1.7899 2.5547</td>
</tr>
<tr>
<td>II</td>
<td>1.9273</td>
<td>6.8710</td>
<td>8.9800</td>
<td>0.0000</td>
<td>1.5065 2.3481</td>
</tr>
<tr>
<td>Data</td>
<td>0.0885</td>
<td>1.0926</td>
<td>0.4900</td>
<td>0.4226</td>
<td>-0.2655 0.4426</td>
</tr>
<tr>
<td>Interview</td>
<td>-0.4260</td>
<td>0.6531</td>
<td>-0.3800</td>
<td>0.7040</td>
<td>-2.6200 1.7679</td>
</tr>
<tr>
<td>Survey</td>
<td>-0.1809</td>
<td>0.8345</td>
<td>-0.3200</td>
<td>0.7500</td>
<td>-1.2933 0.9316</td>
</tr>
<tr>
<td>Document</td>
<td>-0.1427</td>
<td>0.8670</td>
<td>-0.2400</td>
<td>0.8070</td>
<td>-1.2899 1.0045</td>
</tr>
<tr>
<td>No. of observations</td>
<td>542</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1995-2004 (NBR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.2452</td>
<td>0.2879</td>
<td>-11.9900</td>
<td>0.0000</td>
<td>-1.4488 - 1.0416</td>
</tr>
<tr>
<td>NJS</td>
<td>-0.0506</td>
<td>0.9507</td>
<td>-1.1200</td>
<td>0.2630</td>
<td>-0.1391 0.0380</td>
</tr>
<tr>
<td>GCD</td>
<td>0.0000</td>
<td>1.0000</td>
<td>2.4700</td>
<td>0.0140</td>
<td>0.0000 0.0001</td>
</tr>
<tr>
<td>IN</td>
<td>1.5337</td>
<td>4.6351</td>
<td>12.5500</td>
<td>0.0000</td>
<td>1.2941 1.7732</td>
</tr>
<tr>
<td>II</td>
<td>1.1910</td>
<td>3.2905</td>
<td>8.1700</td>
<td>0.0000</td>
<td>0.9054 1.4766</td>
</tr>
<tr>
<td>Data</td>
<td>0.6350</td>
<td>1.8871</td>
<td>7.0300</td>
<td>0.0000</td>
<td>0.4580 0.8121</td>
</tr>
<tr>
<td>Interview</td>
<td>-0.0211</td>
<td>0.9791</td>
<td>-0.0700</td>
<td>0.9450</td>
<td>-0.6197 0.5775</td>
</tr>
<tr>
<td>Survey</td>
<td>1.0831</td>
<td>2.9538</td>
<td>6.5800</td>
<td>0.0000</td>
<td>0.7604 1.4057</td>
</tr>
<tr>
<td>Document</td>
<td>-0.4076</td>
<td>0.6652</td>
<td>-1.1000</td>
<td>0.2700</td>
<td>-1.1323 0.3171</td>
</tr>
<tr>
<td>No. of observations</td>
<td>746</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005-2014 (Poisson)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.8757</td>
<td>0.4166</td>
<td>-13.6600</td>
<td>0.0000</td>
<td>-1.0013 - 0.7501</td>
</tr>
</tbody>
</table>
In the first period, interview, survey and document based research articles have a tendency to have in average fewer authors in the first period (~10-35% fewer co-authors), while the usage of data do not seem to have any influence either positive or negative. This changes for articles using survey and data in the second period, where a survey article have 195% and a data article 100% more co-authors. The coefficient decreases in the third period, but these types of research methods still have a positive effect on the number of co-authors. Articles using interviews have a negative coefficient in the second period, but it increases in the third period, so interview give 35% more co-authors in average in the third period. The coefficient for document analysis articles remains negative for the whole period, and it seems that researchers using this kind of data collection prefers to be single authors. The numerous tasks in relation with data collection and preparation for interview and surveys could be the reason these types of research have more co-authors. Similar, the link between a higher number of co-authors and data could be because it requires some data preparation, but it can also be easier to share data obtain from archives or databases. The model for the whole period show that the results seems robust except for the variable NJS and document.

**Conclusion**

This study shows there have been changes in scholarly communication and collaboration. Danish economics research have experienced a shift from mainly working as a theoretical field to a more empirical field. It finds that even though the field of economics is quantitative based, it has become more diverse in its research approach during the last couple of years. It seems that when researchers conduct quantitative studies with primary data, such as surveys, there are a greater number of authors than when the data sample is from registers or databases. However, articles with interview studies are generally less co-authored and have fewer authors, but this also seems to change in the last period.

This study examines how the number of authors have changed in Danish economics, and how changes in external factors, such as the kind of research methods and approaches applied creates changes in co-authorship. This study finds that to some extent the changes in economics have contributed to the increasing number of authors and co-authored articles. Nevertheless, it also finds that since the increase in the number of authors is occurring across the field of economics, no matter the research methods or approaches. These changes indicate that the research culture is changing. The question is whether researchers nowadays collaborate more or the increases in number of authors is a reflection of changes in how researchers perceives and practices co-authorship.

**References**


Thelwall, M., & Sud, P. (2016). National, disciplinary and temporal variations in the extent to which articles with more authors have more impact: Evidence from a geometric field normalised citation indicator. *Journal of Informetrics, 10*(1), 48-61.

Introduction

This paper describes the forecasting methods used to analyze the future demand for domain names under the ccTLD (country code Top-Level Domain) “.it”. The aim of this analysis was to examine the trend over time of the demand for .it domain names. This is firstly to assess whether we are in a phase of growth and expansion or at a point of saturation. Secondly, this analysis enables us to compare the .it domain name market with those of other ccTLDs (.de, .uk, etc.) or of other gTLDs - general Top-Level Domains (.com, .edu, etc.).

Methods

In order to forecast the trend of demand for .it domain names, the main forecasting methods available in the literature were used. Before applying these forecasting methods, there was an analysis of the trend and seasonality of the series “variation in growth of .it” domain names. This is none other than the difference between registrations and cancellations recorded in the quarter x of year t.

In general, a historical series can be broken down into a seasonal component, a trend and an erratic component. These components are usually estimated using the moving average. However, this method, albeit commonly used in the literature because of its ease of calculation, has some disadvantages. These are associated both with the limited accuracy of the estimates obtained and the loss of data relative to some terms due to the smoothing of the moving average.

On the basis of these considerations, the trend, the seasonality and the erratic component of “.it” domains were estimated using the loess method which, differently from the abovementioned method, gives more accurate estimates (Ricci, 2005). Subsequently, in order to estimate the forecasts of the .it domain names registration, some forecasting methods were applied. These included exponential smoothing methods, such as the Holt-Winters (additive and multiplicative) (Holt, 1957; Winters, 1960), which consider the presence of the trend and seasonality, and the stochastic processes like ARIMA - Auto Regressive Integrated Moving Average models (Box and Jenkins approach) (Box and Jenkins, 1976), in particular the SARIMA (seasonal ARIMA) version. The parameters of the various models were calculated using the maximum likelihood estimation. To assess the accuracy of the forecasts, resulting from the various forecasting methods, a range of indicators were used and analyzed which were able to give a measurement of the forecast error. These included the ME (Mean Error), MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), MPE (Mean Percentage Error), MAPE (Mean Absolute Percentage Error) and MASE (Mean Absolute Scaled Error). The latter indicator, was proposed by Hyndman and Koehler (2006) as an alternative to using percentage errors when comparing forecast accuracy across series on different scales. Furthermore, due to the predictive accuracy of the methods considered, also other indicators were analyzed such as the Akaike Information Criterion (AIC), Corrected Akaike Information Criterion (AICc) and Schwarz Bayesian Information Criterion (BIC). Finally, in order to identify the best forecasting method, also the correlogram of the residuals of the various models were analyzed (ACF-Fuction of autocorrelation and PACF-Fuction of partial autocorrelation). The application of these forecasting methods and statistical indicators was carried out using the “R” statistical environment. The registrations and cancellations of .it domains were extracted in the .it domain names database managed by the Institute for Informatics and Telematics (IIT) of the CNR (National Research Council) in Pisa. The analysis was done taking into account the time period of nine years, from 2008 to 2016. Each year was subdivided into four quarters.

Results

From the analysis of the calculation of the trend and seasonal components, our time series shows a trend (slow growth during 2008, steadying between 2008 and 2009, reaching its maximum in the first half year of 2011, decreasing after 2011 until 2014, renewed growth to subsequently steady again). Furthermore, there is a seasonal component, given that the registrations and cancellations of .it domains were extracted in the .it domain names database managed by the Institute for Informatics and Telematics (IIT) of the CNR (National Research Council) in Pisa. The analysis was done taking into account the time period of nine years, from 2008 to 2016. Each year was subdivided into four quarters.
seasonal component, so as to make the series stationary, due to the seasonality in the demand for domain names. The choice of this model, compared to other ARIMA models, was made based on an analysis of the residuals and their correlograms (ACF and PACF).

Besides the ARIMA model, the other models applied to our series were both the Holt–Winters model with damped trend (Gardner and McKenzie, 1985) and additive seasonal component, and the Holt-Winters model with multiplicative trend and seasonal component. Also in this case the choice of these models instead of others was determined by considering the residuals and their correlograms. Figure 1 shows application of the abovementioned models to the .it domain name time series (registrations - cancellations).

As shown in figure 1, the ARIMA model fits our data better than the Holt-Winters models, above all in the first years (in particular in 2008 the values calculated by the ARIMA model are equal to the observed values).

| Figure 1. Estimates and forecasts of .it domains using Holt-Winters and ARIMA models. |
|---|---|---|---|
| Period | ARIMA | H-W Multiplicative | H-W Additive Damped |
| Q1 2017 | 54,995 | 52,379 | 48,920 |
| Q2 2017 | 16,732 | 17,031 | 26,313 |
| Q3 2017 | 32,776 | 29,968 | 24,596 |
| Q4 2017 | 30,470 | 37,228 | 39,065 |
| Q1 2018 | 54,781 | 54,201 | 48,454 |
| Q2 2018 | 16,613 | 17,624 | 25,941 |
| Q3 2018 | 32,709 | 31,010 | 24,299 |
| Q4 2018 | 30,433 | 38,523 | 38,827 |

Table 1 Forecasts of .it domains using Holt-Winters (H-W) and ARIMA models.

Discussions and Conclusions

The estimates of the parameters of the Holt-Winters damped additive model indicate that the estimate of the level at the current time point is based on the less recent observations (the less recent observations have a greater weight than the more recent ones). On the other hand, for the Holt-Winters multiplicative model the estimate of the level at the current time point is based not only on the less recent observations but also on the most recent ones.

From analysis of the forecasts, the ARIMA model is the one which better fits the historical data compared with the models of exponential smoothing. This conclusion is valid, both when observing the graphs relative to the diagnostics of the residuals and the statistical indicators relative to the accuracy of the estimates. In fact, all the statistical indicators that give a measurement of the error of the estimates (ME, MAE, RMSE, MPE, MAPE and MASE) are lower when applying the ARIMA model, compared to the other two models.

As regards the diagnostics of the residuals, the p values for the Ljung-Box statistic in the various lags are higher for the ARIMA model, indicating that the hypothesis of absence of autocorrelation of the residuals, for all the lags, cannot be rejected. On the contrary, for the Ljung-Box statistic of the Holt-Winters models, the p values are very low compared to the ARIMA model and in some cases the test rejects the hypothesis of uncorrelated residuals. The same applies to other indicators which give a measure of the degree of fit of the models used. AIC, AICc and BIC are much lower for the ARIMA model than for the models of exponential smoothing.

In conclusion, according to the results in Table 1, the .it domain name registrations will be greater than the cancellations in both 2017 and 2018.

In summary, notwithstanding the forecast of a growth in domain names in the 2017-2018 two-year period, the increase is however limited, tending to reach a phase of saturation of the “.it” market.

References


Holt, C.C. (1957). Forecasting Seasonal and Trends by exponential weighted moving averages. Research Memorandum No. 52


On the Applicability of Altmetrics in the Evaluation of Scientific Journals

Zhang Yang¹ Lun Huilian²

¹ zhyang2@mail.sysu.edu.cn
Sun Yat-sen University, Guangzhou (China)

² huilianlun@163.com
Sun Yat-sen University, Guangzhou (China)

Abstract
This study aims to explore the applicability of Altmetrics in the evaluation of scientific journals. Based on a sample of 10,041 journals, it provides a comparison between Altmetrics metrics and citation metrics, taking correlation analysis as the main method. It also discusses the applicable scope of the Altmetrics metrics in journal evaluation from four aspects. Results indicate that the Altmetrics metrics calculated with the Altmetrics data of articles can reflect the influence of journals. Also, these metrics have good discrimination on the academic quality of journals. In most cases, sum metrics of Altmetrics metrics are more stable and are more able to reflect the academic influence of journals. The choice of Altmetrics data sources will affect the journal influence it reflects. Currently Altmetrics is not applicable for specific disciplines or publishing areas. Altmetrics achieves better results in journals in Social Science and Health Science, but does worse in Natural Science especially Computer Science and Mathematics. In addition, the Open Access status of journals will affect the evaluation results obtained by Altmetrics.

Conference Topic
Altmetrics; Indicators; Journals, databases and electronic publications; The application of informetrics on evaluation

Introduction
Journal evaluation is an important application realm of Bibliometrics and Scientometrics with citation analysis as its main approach. However, as the academic environment changed and new publishing ways appeared, the inadequacies of citation databases and citation metrics for scientific evaluation are gradually exposed (Agrawal, 2005; Lange, 2002). Papers in journals are no longer the sole carrier of scientific output, since there are many other alternative options including books, proceedings, reports and those come from the Internet (e.g., websites, blogs, academic communities and social media) (Barjak, 2006; Rowlands et al., 2011; Shema et al., 2012). However, citation databases mainly cover journals rather than other means of scientific knowledge diffusion (Mongeon & Paul-Hus, 2016), and the citations to non-source publications could have a significant effect on rankings (Butler & Visser, 2006). Meanwhile, the Internet provides diverse platforms and approaches for scientific outputs to be published and disseminated, but citation metrics neglected that Internet was an approach for the diffusion of knowledge (Gonzalez-Alcaide et al. 2012). These metrics highly focus on citations instead of other usage. Thus, it gets more difficult for citation analysis to satisfy the requirements of evaluation.

Considering the inadequacies of citation analysis for scientific evaluation and the characteristics of academic communication in the Internet, the concept of Altmetrics was raised. Different from the traditional methods based on citations, Altmetrics emphasizes on the data generated in the online academic communication, especially the data from social networks. Altmetrics studies the application of such data, and provides a method of measuring impacts from different aspects, which can be a useful supplement to the traditional evaluation approaches. Now related theoretical studies and practices are being performed. In studies on Article-level Metrics, some metrics of Altmetrics already have plenty empirical studies and have put into use. Therefore,
this paper aims to study the extent to which Altmetrics is applicable in the evaluation of scientific journals, for feasible methods with online data and better evaluation results.

**Literature Review**

Studies on Altmetrics began since 2008. Li, Thelwall and Giustini (2011) adopted CiteULike and Mendeley as the data sources, and concluded that online reference manager tools could be used to measure the academic influence of articles as supplementary data source. Priem, Piwowar and Hemminger (2012) took articles published by PLoS as sample and presented a study on the correlation between Altmetrics and citation metrics. They found that the influences reflected by these two metrics are different, so it cannot describe the academic influence completely with just one of them. Costas, Zahedi and Wouters (2015) had similar conclusion on metrics from Altmetric.com and citation metrics, and they found that social media provided low value of Altmetrics metrics, and disciplines including social science, humanities, medicine and life science got high value of metrics.

Studies focusing on specific Altmetrics metrics present their applicability and characteristics. The number of readers in Mendeley and the number of tweets arose plenty of interest. Results of the correlation analysis showed that the correlation coefficient between citation number and reader number was stable at 0.6 (Maflahi & Thelwall, 2016). The reader number of Mendeley could be used to measure the influence of articles published in different time. The data distribution of Mendeley metrics was most similar to that of citation metrics (Costas, Zahedi & Wouters, 2015), while the citation behavior was quite different from the behavior on Twitter (de Winter, 2015). However, using the data from both of Mendeley and Twitter would cause differences between disciplines. The reader number of Mendeley was more correlated to citation metrics in social science than in the humanities (Mohammadi & Thelwall, 2014). Twitter paid more attention to the medicine and the social science while the humanities, the nature science and the engineering science achieved lower value of metrics from social media (Costas, Zahedi & Wouters, 2015).

Further studies made it more comprehensively. Thelwall et al. (2013) selected 208,739 articles with at least one Altmetrics data record from PubMed to conduct a study on 11 Altmetrics metrics. They found that in the medicine and the biological science, 6 Altmetrics metrics have significant correlation with citations including, Google+ posts, Tweets, Facebook wall posts, research highlights, blogs, mainstream media and forums, while there was no correlation found between citations and LinkedIn, Pinners, Q&A and Reddits.

Plenty of empirical studies on Altmetrics to date are about Article-Level Metrics. Apart from this, there are studies considering Altmetrics on journal level. Bollen et al. (2011) compared a set of network metrics generated from citation and download data with ISI IF. Results indicate that, although social network metrics and ISI IF rankings deviate moderately for citation-based journal networks, they differ considerably for journal networks derived from download data. Schlögl et al. (2014) chose two journals on information systems as sample and made the correlation analysis between the numbers of downloads in Science Direct, the citations in Scopus and the number of readers in Mendeley. They found that these two Altmetrics metrics had correlation with citations but there were also some differences existed. Sotudeh, Mazarei & Mirzabeigi (2015) studied the relationship between the bookmarks in CiteULike and the traditional impact metrics on the level of journals and scholars. They concluded that the bookmark data could be used as a supplement to the traditional metrics, but cannot replace them. Fan, Peng & Shen (2015) presented an empirical analysis taking Scientometrics as the sample, and concluded that when applying Altmetrics to the journal evaluation, researchers should differentiate data from different sources, increase analysis on the evaluation time and identify the manipulation of attention.
Given that studies on Altmetrics have a good start on journal level, this study aims to present a further exploration. Since most of studies on Altmetrics adopted correlation analysis as the main method, and the results are helpful to find out the similarity and difference between metrics, this study takes correlation analysis as the main method to explore the relationship between citation metrics and Altmetrics on the journal level. Also, we will discuss and summarize the applicability of Altmetrics in the evaluation of journals from various perspectives.

Methodology

Sample
The sample of this study is selected from the Scopus Source List (November 2015). There are four source types in this list, and only Journal and Trade Journal were adopted by this study. For identification, only unique records with Print-ISSN or Electronic-ISSN or both of them were included. Records which failed to be obtained SJR2014 or Altmetrics data were excluded. These resulted in 10,041 journals as the sample of this study. In the following section, the publication types of journals are identified according to their Print-ISSN and E-ISSN. The disciplines are identified according to the All Science Journals Classification (ASJC) code provided by Scopus, and some interdisciplinary journals have more than one ASJC code. Therefore, the numbers of journals in Life Sciences, Social Sciences, Physical Sciences and Health Sciences are 2,386, 2,699, 2,032 and 4,257.

Data
The Scopus Source List (November 2015) was downloaded for the sample of this study. The information of sample journals was extracted from it including Source Title, P-ISSN, E-ISSN, Coverage, Open Access Status, Publisher’s Country and ASJC. The citation data adopted by this study comes from SCImago Journal Rankings (2014). We matched journals by ISSN and collected their SJR, H-index, total number of documents published in 2014 (Total Docs. (2014)) and from 2011 to 2013 (Total Docs. (3 years)), total number of references cited by documents published in 2014 (Total Refs.), total number of citations in 2014 of documents published from 2011 to 2013 (Total Cites (3 years)), total number of citable documents from 2011 to 2013 (Citable Docs. (3 years)), average number of citations per document in the last 2 years (Cites / Doc. (2 years)) and average number of references per document (Ref. / Doc.).

Altmetric Explorer was used to collect Altmetrics data. With its filters “In these journals” and “Mentioned in the past”, we obtained data of papers published in sample journals and mentioned in the past 1 year. As a result, Altmetrics data of 1,149,369 articles from 10,041 journals was collected including total number of articles, usage of various sources, number of readers and Altmetrics score.

Metrics
Metrics of journals are divided into 3 groups according to their data sources, shown as Table 1, Table 2 and Table 3. Altmetrics Comprehensive Metrics group includes metrics calculated with data from various sources. Altmetrics Source Metrics group includes metrics coming from single data source. Citation Metrics group includes metrics related to citation data.

Procedures
In the following sections, firstly we calculate the correlation coefficients of citation metrics and Altmetrics metrics on journal basis, and provide a detailed comparison between different metrics and discuss the applicability of Altmetrics metrics. Secondly with coefficient of variation, we make a discussion on the discrimination of Altmetrics metrics. Thirdly we present
a comparison and discuss the applicable scope of Altmetrics metrics from four aspects: discipline, publishing area, publishing type and Open Access status.

### Table 1. Altmetrics comprehensive metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Articles Sum</td>
<td>Number of articles, which had been mentioned in the last year, captured by Altmetric.com of a journal</td>
</tr>
<tr>
<td>Accounts Sum</td>
<td>Number of accounts which had mentioned the articles in a journal</td>
</tr>
<tr>
<td>Post Sum</td>
<td>Number of posts which had mentioned the articles in a journal</td>
</tr>
<tr>
<td>Readers Sum</td>
<td>Number of readers who had used the articles in a journal</td>
</tr>
<tr>
<td>Readers Mean</td>
<td>Number of Readers Sum divided by Articles Sum of a journal</td>
</tr>
<tr>
<td>Readers Median</td>
<td>Median of readers sum of articles in a journal</td>
</tr>
<tr>
<td>1-year Altmetrics Scores Sum</td>
<td>Sum of Altmetrics Scores in the last year of articles, which had been mentioned in the last year, in a journal</td>
</tr>
<tr>
<td>1-year Altmetrics Scores Mean</td>
<td>Number of 1-year Altmetrics Scores divided by Articles Sum</td>
</tr>
<tr>
<td>1-year Altmetrics Scores Median</td>
<td>Median of Altmetrics Scores in the last year of articles in a journal</td>
</tr>
<tr>
<td>Altmetrics Scores Sum</td>
<td>Sum of Altmetrics Scores of articles, which had been mentioned in the last year, in a journal</td>
</tr>
<tr>
<td>Altmetrics Scores Mean</td>
<td>Number of Altmetrics Scores divided by Articles Sum</td>
</tr>
<tr>
<td>Altmetrics Scores Median</td>
<td>Median of Altmetrics Scores of articles in a journal</td>
</tr>
</tbody>
</table>

### Table 2. Altmetrics Source metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Src_fbwalls</td>
<td>Number of Facebook Pages</td>
</tr>
<tr>
<td>Src_feeds</td>
<td>Number of Blogs</td>
</tr>
<tr>
<td>Src_gplus</td>
<td>Number of Google+ users</td>
</tr>
<tr>
<td>Src_linkedin</td>
<td>Number of LinkedIn users</td>
</tr>
<tr>
<td>Src_msm</td>
<td>Number of news outlets</td>
</tr>
<tr>
<td>Src_peer_review_sites</td>
<td>Number of peer review sites</td>
</tr>
<tr>
<td>Src_pinters</td>
<td>Number of Pinterest posts</td>
</tr>
<tr>
<td>Src_policies</td>
<td>Number of policy documents</td>
</tr>
<tr>
<td>Src_qna</td>
<td>Number of Q&amp;A site users</td>
</tr>
<tr>
<td>Src_rds</td>
<td>Number of Redditors</td>
</tr>
<tr>
<td>Src_rh</td>
<td>Number of F1000 reviews</td>
</tr>
<tr>
<td>Src_tweeters</td>
<td>Number of Tweeters</td>
</tr>
<tr>
<td>Src_videos</td>
<td>Number of videos</td>
</tr>
<tr>
<td>Src_weibo</td>
<td>Number of Weibo users</td>
</tr>
<tr>
<td>Src_wikipedia</td>
<td>Number of Wikipedia Pages</td>
</tr>
</tbody>
</table>

### Table 3. Citation metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SJR2014</td>
<td>SCImago Journal Rank in 2014</td>
</tr>
<tr>
<td>H-index</td>
<td>H-index in 2014</td>
</tr>
<tr>
<td>Total Docs. (2014)</td>
<td>Number of documents published in 2014</td>
</tr>
<tr>
<td>Total Docs. (3 years)</td>
<td>Number of documents published from 2011 to 2013</td>
</tr>
<tr>
<td>Total Cites (3 years)</td>
<td>Number of citations received by a journal in 2014 to documents published from 2011 to 2013</td>
</tr>
<tr>
<td>Cites / Doc. (2 years)</td>
<td>Number of citations received by a journal in 2014 to documents published from 2012 to 2013, divided by the number of documents indexed in Scopus published from 2012-2013</td>
</tr>
</tbody>
</table>

**Findings**

**Altmetrics comprehensive metrics**

Table 4 shows the correlation coefficients between 12 Altmetrics comprehensive metrics and 6 citation metrics. Most of Altmetrics metrics have positive correlation with citation metrics...
except for 1-year Altmetrics Scores Median and Altmetrics Scores Median. With low coefficients, both of the two median metrics cannot reflect the academic influence or the quantity of articles. Differently, Readers Median has moderate positive correlation with citation metrics. It can reflect recent influence rather than accumulated influence. However, compared with Readers Mean and Readers Sum, its ability to reflect journal influence is inferior. Therefore, median metrics are not applicable in the evaluation of journals.

Table 4. Correlation coefficients between Altmetrics comprehensive metrics and citation metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>SJR2014</th>
<th>Cites / Doc. (2 years)</th>
<th>Total Cites (3 years)</th>
<th>H-index</th>
<th>Total Docs. (2014)</th>
<th>Total Docs. (3 years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Articles Sum</td>
<td>0.672</td>
<td>0.711</td>
<td>0.799</td>
<td>0.730</td>
<td>0.706</td>
<td>0.683</td>
</tr>
<tr>
<td>Accounts Sum</td>
<td>0.678</td>
<td>0.709</td>
<td>0.741</td>
<td>0.680</td>
<td>0.602</td>
<td>0.571</td>
</tr>
<tr>
<td>Post Sum</td>
<td>0.671</td>
<td>0.702</td>
<td>0.733</td>
<td>0.672</td>
<td>0.595</td>
<td>0.564</td>
</tr>
<tr>
<td>Readers Mean</td>
<td>0.700</td>
<td>0.661</td>
<td>0.554</td>
<td>0.605</td>
<td>0.255</td>
<td>0.230</td>
</tr>
<tr>
<td>Readers Median</td>
<td>0.638</td>
<td>0.609</td>
<td>0.490</td>
<td>0.526</td>
<td>0.203</td>
<td>0.172</td>
</tr>
<tr>
<td>Readers Sum</td>
<td>0.811</td>
<td>0.816</td>
<td>0.814</td>
<td>0.798</td>
<td>0.589</td>
<td>0.560</td>
</tr>
<tr>
<td>1-year Altmetrics Scores Mean</td>
<td>0.404</td>
<td>0.361</td>
<td>0.262</td>
<td>0.303</td>
<td>0.069</td>
<td>0.042</td>
</tr>
<tr>
<td>1-year Altmetrics Scores Median</td>
<td>0.176</td>
<td>0.097</td>
<td>-0.017</td>
<td>0.068</td>
<td>-0.165</td>
<td>-0.185</td>
</tr>
<tr>
<td>1-year Altmetrics Scores Sum</td>
<td>0.715</td>
<td>0.731</td>
<td>0.760</td>
<td>0.719</td>
<td>0.604</td>
<td>0.572</td>
</tr>
<tr>
<td>Altmetrics Scores Mean</td>
<td>0.439</td>
<td>0.402</td>
<td>0.302</td>
<td>0.338</td>
<td>0.093</td>
<td>0.068</td>
</tr>
<tr>
<td>Altmetrics Scores Median</td>
<td>0.228</td>
<td>0.156</td>
<td>0.028</td>
<td>0.110</td>
<td>-0.146</td>
<td>-0.166</td>
</tr>
<tr>
<td>Altmetrics Scores Sum</td>
<td>0.713</td>
<td>0.729</td>
<td>0.751</td>
<td>0.711</td>
<td>0.585</td>
<td>0.555</td>
</tr>
</tbody>
</table>

Comparing Readers Mean, 1-year Altmetrics Scores Mean and Altmetrics Scores Mean with their sum metrics, the coefficients of sum metrics are all higher than 0.7, obviously higher than mean metrics. It shows that sum metrics are more able to reflect journals’ academic influence, while they are also indicating other influence as Altmetrics metrics.

Articles Sum is the number of articles, which had been mentioned in the last year, captured by Altmetric.com of a journal. It roughly shows the online usage of the journal. The correlation coefficients of Articles Sum with Total Docs. (2014) and Total Docs. (3 years) are 0.706 and 0.683, indicating high correlations. This shows that the visibility of a journal is related to the scale and recent status of the journal, but there exist some factors affecting the online usage of journals. Except for Articles Sum, other Altmetrics metrics have higher coefficients with Total Docs. (2014) than coefficients with Total Docs. (3 years). It once again shows that Altmetrics metrics are more sensitive to the recent status of journals.

In order to know further about the window of Altmetrics Scores, we compare 1-year Altmetrics Scores with Altmetrics Scores. Both the mean and the sum in 2 windows have very close coefficients with citation metrics. Therefore, the window makes little effects on the applicability of Altmetrics Scores in the evaluation of journals.

Comparing coefficients between Altmetrics sum metrics and SJR2014, all of them show moderate to high positive correlations. Readers Sum achieves the highest coefficient at 0.811, while Accounts Sum gets the lowest, and Altmetrics Scores Sum is between two of them. It shows that Readers Sum can get the evaluation results closest to that of citation metrics. As the data used by a metric contains data from more social media sources, the metric tends to reflect more complex influence of journals. As a result, when Altmetrics is applied to evaluation, online reference manager tools and social media should be differentiated as two different data...
satisfactory. It is meaningful to consider how to combine them for evaluation to get a more comprehensive result.

Table 5 shows the correlation coefficients between 9 Altmetrics comprehensive metrics. Obviously, 6 sum metrics have high correlation between them with coefficient higher than 0.85. 3 mean metrics have weak positive correlation with Articles Sum, which shows that Altmetrics metrics are also affected by the scale of journals. Therefore, mean metrics can help to reduce such effects.

<table>
<thead>
<tr>
<th></th>
<th>Articles Sum</th>
<th>Accounts Sum</th>
<th>Post Sum</th>
<th>Readers Mean</th>
<th>Readers Sum</th>
<th>1-year Altmetrics Scores</th>
<th>1-year Altmetrics Scores</th>
<th>Altmetrics Scores</th>
<th>Altmetrics Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Articles Sum</td>
<td>1.000</td>
<td>0.944</td>
<td>0.937</td>
<td>0.442</td>
<td>0.874</td>
<td>0.267</td>
<td>0.925</td>
<td>0.320</td>
<td>0.913</td>
</tr>
<tr>
<td>Accounts Sum</td>
<td>0.944</td>
<td>1.000</td>
<td>0.996</td>
<td>0.525</td>
<td>0.886</td>
<td>0.458</td>
<td>0.962</td>
<td>0.538</td>
<td>0.973</td>
</tr>
<tr>
<td>Post Sum</td>
<td>0.937</td>
<td>0.996</td>
<td>1.000</td>
<td>0.524</td>
<td>0.881</td>
<td>0.457</td>
<td>0.955</td>
<td>0.539</td>
<td>0.967</td>
</tr>
<tr>
<td>Readers Mean</td>
<td>0.442</td>
<td>0.525</td>
<td>0.524</td>
<td>1.000</td>
<td>0.801</td>
<td>0.498</td>
<td>0.556</td>
<td>0.562</td>
<td>0.582</td>
</tr>
<tr>
<td>Readers Sum</td>
<td>0.874</td>
<td>0.886</td>
<td>0.881</td>
<td>0.801</td>
<td>1.000</td>
<td>0.434</td>
<td>0.892</td>
<td>0.502</td>
<td>0.899</td>
</tr>
<tr>
<td>1-year Altmetrics Scores Mean</td>
<td>0.267</td>
<td>0.458</td>
<td>0.457</td>
<td>0.498</td>
<td>0.434</td>
<td>1.000</td>
<td>0.584</td>
<td>0.941</td>
<td>0.585</td>
</tr>
<tr>
<td>1-year Altmetrics Scores Sum</td>
<td>0.925</td>
<td>0.962</td>
<td>0.955</td>
<td>0.556</td>
<td>0.892</td>
<td>0.584</td>
<td>1.000</td>
<td>0.612</td>
<td>0.990</td>
</tr>
<tr>
<td>Altmetrics Scores Sum</td>
<td>0.320</td>
<td>0.538</td>
<td>0.539</td>
<td>0.562</td>
<td>0.502</td>
<td>0.941</td>
<td>0.612</td>
<td>1.000</td>
<td>0.653</td>
</tr>
<tr>
<td>Altmetrics Scores Mean</td>
<td>0.913</td>
<td>0.973</td>
<td>0.967</td>
<td>0.582</td>
<td>0.899</td>
<td>0.585</td>
<td>0.990</td>
<td>0.653</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Accounts Sum is the number of users of a journal in social media, while Readers Sum is the number of users of a journal in reference manager tools. The correlation coefficient between them is 0.886, showing a high correlation. It indicates that the numbers of users of a journal in different online platforms are similar. But since they are platforms with different types of content, reference manager tools and social media are quite different when being used as Altmetrics data sources. Comparing the correlation of Altmetrics Scores Sum with Accounts Sum and Readers Sum, the coefficient of Accounts Sum is significantly higher than that of Readers Sum. This is due to the data sources of Altmetrics Scores. Since reference manager tools only take up a small part of the online platform, social media will have greater effects on Altmetrics Scores.

**Altmetrics source metrics**

Table 6 shows the correlation coefficients between 15 Altmetrics source metrics and 6 citation metrics. All Altmetrics source metrics have positive correlation with citation metrics with coefficients from 0.13 to 0.72. Src_tweeters, Src_msm and Src_fbwalls have high coefficient, while Src_linkedin and Src_pinners have the lowest coefficient compared with other sources. Since LinkedIn is a business and employment-oriented social networking platform and Pinterest is a photo sharing website, both of them focus on a specific field which is not related to scientific research. Online platforms with more comprehensive interests turn out to be more suitable in scientific evaluation.

The coefficients between Altmetrics source metrics and Total Docs. (2014), and the coefficients between Altmetrics source metrics and Total Docs. (3 years) are calculated. Figure 1 shows the difference between them. Sources with higher coefficients of Total Docs. (3 years) includes LinkedIn, Pinterest, policy documents, Q&A sites, videos and Wikipedia. Except for LinkedIn and Pinterest, which focus on a specific field, other 4 sources have less social activities or are
updated more slowly. Sources with active social connection such as Twitter and Facebook have greater coefficient differences. Also, platforms require quick update have obvious differences such as news outlets and blogs. In other words, the short-time change of journals can be reflected more easily in the platforms with faster update speed. Therefore, a feasible combination of different Altmetrics sources will help to make a more comprehensive evaluation of journals’ impact.

Table 6. Correlation coefficients between Altmetrics source metrics and citation metrics

<table>
<thead>
<tr>
<th></th>
<th>SJ2014</th>
<th>Cites / Doc. (2 years)</th>
<th>Total Cites (3 years)</th>
<th>H-index</th>
<th>Total Docs. (2014)</th>
<th>Total Docs. (3 years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Src_fbwalls</td>
<td>0.574</td>
<td>0.634</td>
<td>0.678</td>
<td>0.602</td>
<td>0.570</td>
<td>0.550</td>
</tr>
<tr>
<td>Src_feeds</td>
<td>0.663</td>
<td>0.666</td>
<td>0.677</td>
<td>0.663</td>
<td>0.501</td>
<td>0.485</td>
</tr>
<tr>
<td>Src_gplus</td>
<td>0.547</td>
<td>0.589</td>
<td>0.606</td>
<td>0.546</td>
<td>0.479</td>
<td>0.462</td>
</tr>
<tr>
<td>Src_linkedin</td>
<td>0.150</td>
<td>0.159</td>
<td>0.161</td>
<td>0.149</td>
<td>0.131</td>
<td>0.134</td>
</tr>
<tr>
<td>Src_msm</td>
<td>0.655</td>
<td>0.685</td>
<td>0.700</td>
<td>0.668</td>
<td>0.535</td>
<td>0.516</td>
</tr>
<tr>
<td>Src_peer_review_sites</td>
<td>0.356</td>
<td>0.395</td>
<td>0.438</td>
<td>0.382</td>
<td>0.388</td>
<td>0.381</td>
</tr>
<tr>
<td>Src_pinners</td>
<td>0.157</td>
<td>0.170</td>
<td>0.172</td>
<td>0.146</td>
<td>0.136</td>
<td>0.137</td>
</tr>
<tr>
<td>Src_policies</td>
<td>0.374</td>
<td>0.385</td>
<td>0.427</td>
<td>0.439</td>
<td>0.353</td>
<td>0.355</td>
</tr>
<tr>
<td>Src_qna</td>
<td>0.365</td>
<td>0.361</td>
<td>0.396</td>
<td>0.393</td>
<td>0.313</td>
<td>0.323</td>
</tr>
<tr>
<td>Src_rds</td>
<td>0.437</td>
<td>0.479</td>
<td>0.497</td>
<td>0.447</td>
<td>0.394</td>
<td>0.388</td>
</tr>
<tr>
<td>Src_rh</td>
<td>0.420</td>
<td>0.465</td>
<td>0.486</td>
<td>0.445</td>
<td>0.410</td>
<td>0.409</td>
</tr>
<tr>
<td>Src_tweeters</td>
<td>0.657</td>
<td>0.689</td>
<td>0.717</td>
<td>0.653</td>
<td>0.588</td>
<td>0.553</td>
</tr>
<tr>
<td>Src_videos</td>
<td>0.360</td>
<td>0.394</td>
<td>0.421</td>
<td>0.400</td>
<td>0.341</td>
<td>0.345</td>
</tr>
<tr>
<td>Src_weibo</td>
<td>0.326</td>
<td>0.330</td>
<td>0.335</td>
<td>0.309</td>
<td>0.273</td>
<td>0.264</td>
</tr>
<tr>
<td>Src_wikipedia</td>
<td>0.596</td>
<td>0.588</td>
<td>0.631</td>
<td>0.682</td>
<td>0.474</td>
<td>0.477</td>
</tr>
</tbody>
</table>

Figure 1. Coefficient difference of Altmetrics source metrics between Total Docs. (2014) and Total Docs. (3 years)

Correlation coefficients between Altmetrics source metrics are calculated. All of them have positive correlations. Similar as above, Src_linkedin and Src_pinners show great differences with other metrics. Src_policies, Src_peer_review_sites, Src_qna and Src_weibo have weak correlation with other metrics. It shows that the similarity between these sources depends more on the characteristics of themselves.
Table 7 shows the correlation coefficients between Altmetrics comprehensive metrics and source metrics. Src_fbwalls, Src_feeds, Src_msm and Src_tweeters have the strongest correlation with Altmetrics comprehensive metrics with coefficients higher than 0.75. Since Twitter achieves the highest coefficient, it can obtain the evaluation results closest to that of Altmetrics comprehensive metrics when we adopt only one Altmetrics source.

Table 7. Correlation coefficients between Altmetrics comprehensive metrics and source metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Articles Scores Mean</th>
<th>Articles Scores Sum</th>
<th>Accounts Scores Mean</th>
<th>Accounts Scores Sum</th>
<th>Post Scores Mean</th>
<th>Post Scores Sum</th>
<th>Readers Scores Mean</th>
<th>Readers Scores Sum</th>
<th>1-year Altmetrics Scores Mean</th>
<th>1-year Altmetrics Scores Sum</th>
<th>3-year Altmetrics Scores Mean</th>
<th>3-year Altmetrics Scores Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Src_fbwalls</td>
<td>0.847</td>
<td>0.892</td>
<td>0.887</td>
<td>0.455</td>
<td>0.787</td>
<td>0.348</td>
<td>0.836</td>
<td>0.436</td>
<td>0.852</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Src_feeds</td>
<td>0.772</td>
<td>0.818</td>
<td>0.810</td>
<td>0.580</td>
<td>0.814</td>
<td>0.575</td>
<td>0.865</td>
<td>0.639</td>
<td>0.881</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Src_gplus</td>
<td>0.691</td>
<td>0.750</td>
<td>0.748</td>
<td>0.477</td>
<td>0.710</td>
<td>0.405</td>
<td>0.729</td>
<td>0.485</td>
<td>0.751</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Src_linkedin</td>
<td>0.177</td>
<td>0.187</td>
<td>0.187</td>
<td>0.142</td>
<td>0.185</td>
<td>0.139</td>
<td>0.185</td>
<td>0.153</td>
<td>0.189</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Src_msm</td>
<td>0.762</td>
<td>0.800</td>
<td>0.793</td>
<td>0.533</td>
<td>0.782</td>
<td>0.616</td>
<td>0.875</td>
<td>0.651</td>
<td>0.878</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Src_peer_review_sites</td>
<td>0.442</td>
<td>0.428</td>
<td>0.427</td>
<td>0.246</td>
<td>0.423</td>
<td>0.183</td>
<td>0.425</td>
<td>0.203</td>
<td>0.423</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Src_pinners</td>
<td>0.192</td>
<td>0.212</td>
<td>0.211</td>
<td>0.173</td>
<td>0.210</td>
<td>0.156</td>
<td>0.204</td>
<td>0.181</td>
<td>0.213</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Src_policies</td>
<td>0.479</td>
<td>0.479</td>
<td>0.480</td>
<td>0.303</td>
<td>0.479</td>
<td>0.250</td>
<td>0.484</td>
<td>0.284</td>
<td>0.487</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Src_qna</td>
<td>0.377</td>
<td>0.378</td>
<td>0.376</td>
<td>0.317</td>
<td>0.418</td>
<td>0.228</td>
<td>0.385</td>
<td>0.268</td>
<td>0.395</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Src_rds</td>
<td>0.577</td>
<td>0.624</td>
<td>0.621</td>
<td>0.367</td>
<td>0.574</td>
<td>0.348</td>
<td>0.605</td>
<td>0.418</td>
<td>0.626</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Src_rdts</td>
<td>0.507</td>
<td>0.494</td>
<td>0.491</td>
<td>0.266</td>
<td>0.477</td>
<td>0.225</td>
<td>0.490</td>
<td>0.246</td>
<td>0.487</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Src_rh</td>
<td>0.931</td>
<td>0.989</td>
<td>0.986</td>
<td>0.507</td>
<td>0.868</td>
<td>0.432</td>
<td>0.941</td>
<td>0.510</td>
<td>0.951</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Src_tweeters</td>
<td>0.464</td>
<td>0.490</td>
<td>0.489</td>
<td>0.325</td>
<td>0.475</td>
<td>0.300</td>
<td>0.483</td>
<td>0.356</td>
<td>0.499</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Src_videos</td>
<td>0.370</td>
<td>0.393</td>
<td>0.391</td>
<td>0.279</td>
<td>0.389</td>
<td>0.283</td>
<td>0.393</td>
<td>0.313</td>
<td>0.402</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Src_weibo</td>
<td>0.683</td>
<td>0.661</td>
<td>0.653</td>
<td>0.501</td>
<td>0.717</td>
<td>0.379</td>
<td>0.709</td>
<td>0.411</td>
<td>0.708</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. Coefficient of variation of metrics
Discrimination of Altmetrics metrics

In order to get further understanding of the discrimination of Altmetrics metrics, Figure 2 shows the coefficient of variation. Compared with citation metrics, except for Articles Sum and Altmetrics Scores Mean, all coefficient of variation of Altmetrics metrics are significantly higher than that of citation metrics. It shows that using Altmetrics metrics to evaluate the influence of journals can obtain a result with obvious difference. Altmetrics metrics make good discrimination in the influence level of journals.

The coefficients of variation of Altmetrics comprehensive metrics are from 129% to 880%, lower than source metrics but higher than citation metrics. It shows that Altmetrics comprehensive metrics have better discrimination than citation metrics. Articles Sum, Readers Mean and Altmetrics Scores Mean have relatively low coefficients. Using Altmetrics comprehensive metrics for the evaluation of journals will make the difference between journals more obvious.

All of Altmetrics source metrics have high coefficients of variation, from 600% to 2300%. Since Altmetrics source metrics directly come from the data of different sources, it means different journals might have different influence in each online platforms. There are great difference between the raw data from Altmetrics sources, so such data has to be processed scientifically for better application.

Applicable scope of Altmetrics metrics

For a deeper understanding of the applicable scope of Altmetrics metrics, we provide an analysis from four aspects: discipline, publishing area, publishing type and Open Access status.

Figure 3. Correlation coefficients between SJR2014 and Altmetrics metrics in disciplines

Figure 3 shows the correlation coefficients between SJR2014 and 5 Altmetrics metrics across disciplines. Obviously, the coefficients in Natural Science are relatively low, indicating that using Altmetrics for the evaluation of natural science journals will have great difference with traditional evaluation results. In these four disciplines, Altmetrics is more suitable for the journal evaluation in Social Science and Health Science. Readers Sum can help to obtain results closer to that of citation metrics, while Altmetrics Scores Mean gets the most different one.

Figure 4 shows the correlation coefficients between SJR2014 and Altmetrics metrics across publishing areas. There are 81 countries or areas in our sample, and Figure 4 shows the top 20 with most number of journals published. Except for the Altmetrics Scores Mean in United Arab Emirates, all Altmetrics metrics have positive correlation with SJR2014. The applicability of sum metrics are stable in journals of different areas, while the applicability of mean metrics
obviously varies. Journals published in India, Brazil, China, Russia and United Arab Emirates are not suitable to be evaluated with Altmetrics. Meanwhile, the mean metrics of Altmetrics are not applicable for the evaluation of journals in New Zealand and South Korea.

Table 8 shows the correlation coefficients between SJR2014 and Altmetrics metrics in 3 different publishing types. The correlation coefficient achieves the highest for printed journals and the lowest for electronic journals. It shows that if a journal is exposed in the Internet more, the evaluation results of Altmetrics will have greater difference with that of citation metrics. Also, the component of influence of such journals will be more complicated.

Table 8. Correlation coefficients between SJR2014 and Altmetrics metrics in publishing types

<table>
<thead>
<tr>
<th>Publishing type*</th>
<th>SJR2014</th>
<th>Cites / Doc. (2 years)</th>
<th>Total Cites (3 years)</th>
<th>H-index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readers Mean P</td>
<td>0.722</td>
<td>0.693</td>
<td>0.586</td>
<td>0.625</td>
</tr>
<tr>
<td>E</td>
<td>0.447</td>
<td>0.437</td>
<td>0.240</td>
<td>0.214</td>
</tr>
<tr>
<td>PE</td>
<td>0.667</td>
<td>0.609</td>
<td>0.501</td>
<td>0.580</td>
</tr>
<tr>
<td>Readers Sum P</td>
<td>0.810</td>
<td>0.816</td>
<td>0.811</td>
<td>0.785</td>
</tr>
<tr>
<td>E</td>
<td>0.635</td>
<td>0.621</td>
<td>0.525</td>
<td>0.379</td>
</tr>
<tr>
<td>PE</td>
<td>0.810</td>
<td>0.815</td>
<td>0.816</td>
<td>0.823</td>
</tr>
<tr>
<td>Altmetrics P</td>
<td>0.444</td>
<td>0.412</td>
<td>0.304</td>
<td>0.339</td>
</tr>
<tr>
<td>Scores Mean E</td>
<td>0.265</td>
<td>0.208</td>
<td>0.093</td>
<td>0.061</td>
</tr>
<tr>
<td>PE</td>
<td>0.425</td>
<td>0.380</td>
<td>0.290</td>
<td>0.338</td>
</tr>
<tr>
<td>Altmetrics P</td>
<td>0.704</td>
<td>0.717</td>
<td>0.735</td>
<td>0.690</td>
</tr>
<tr>
<td>Scores Sum E</td>
<td>0.561</td>
<td>0.508</td>
<td>0.507</td>
<td>0.306</td>
</tr>
<tr>
<td>PE</td>
<td>0.723</td>
<td>0.747</td>
<td>0.776</td>
<td>0.751</td>
</tr>
<tr>
<td>Accounts Sum P</td>
<td>0.665</td>
<td>0.692</td>
<td>0.723</td>
<td>0.656</td>
</tr>
<tr>
<td>E</td>
<td>0.522</td>
<td>0.472</td>
<td>0.513</td>
<td>0.305</td>
</tr>
<tr>
<td>PE</td>
<td>0.691</td>
<td>0.732</td>
<td>0.771</td>
<td>0.722</td>
</tr>
</tbody>
</table>

* P: Printed journal; E: Electronic journal; PE: Printed journals with electronic version

Open access status is one of the most influential factor affecting journals’ impact in the Internet. Table 9 shows the correlation coefficients between SJR2014 and Altmetrics metrics in Open Access status. It shows that readers metrics are more able to reflect the academic influence of non-OA journals, while Altmetrics Scores and Accounts Sum do better for OA journals.
Table 9. Correlation coefficients between SJR2014 and Altmetrics metrics in Open Access status

<table>
<thead>
<tr>
<th></th>
<th>Open Access Status</th>
<th>SJR2014</th>
<th>Cites / Doc. (2 years)</th>
<th>Total Cites (3 years)</th>
<th>H-index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OA</td>
<td>0.648</td>
<td>0.624</td>
<td>0.480</td>
<td>0.510</td>
</tr>
<tr>
<td></td>
<td>Not OA</td>
<td>0.710</td>
<td>0.668</td>
<td>0.567</td>
<td>0.628</td>
</tr>
<tr>
<td>Readers Mean</td>
<td>OA</td>
<td>0.806</td>
<td>0.798</td>
<td>0.794</td>
<td>0.727</td>
</tr>
<tr>
<td></td>
<td>Not OA</td>
<td>0.812</td>
<td>0.819</td>
<td>0.817</td>
<td>0.820</td>
</tr>
<tr>
<td>Readers Sum</td>
<td>OA</td>
<td>0.508</td>
<td>0.502</td>
<td>0.339</td>
<td>0.339</td>
</tr>
<tr>
<td></td>
<td>Not OA</td>
<td>0.426</td>
<td>0.384</td>
<td>0.295</td>
<td>0.338</td>
</tr>
<tr>
<td>Altmetrics</td>
<td>OA</td>
<td>0.755</td>
<td>0.758</td>
<td>0.754</td>
<td>0.661</td>
</tr>
<tr>
<td></td>
<td>Not OA</td>
<td>0.706</td>
<td>0.725</td>
<td>0.751</td>
<td>0.728</td>
</tr>
<tr>
<td>Altmetrics Scores Mean</td>
<td>OA</td>
<td>0.729</td>
<td>0.742</td>
<td>0.746</td>
<td>0.635</td>
</tr>
<tr>
<td></td>
<td>Not OA</td>
<td>0.670</td>
<td>0.703</td>
<td>0.741</td>
<td>0.699</td>
</tr>
<tr>
<td>Altmetrics Scores Sum</td>
<td>OA</td>
<td>0.755</td>
<td>0.758</td>
<td>0.754</td>
<td>0.661</td>
</tr>
<tr>
<td></td>
<td>Not OA</td>
<td>0.706</td>
<td>0.725</td>
<td>0.751</td>
<td>0.728</td>
</tr>
<tr>
<td>Accounts Sum</td>
<td>OA</td>
<td>0.729</td>
<td>0.742</td>
<td>0.746</td>
<td>0.635</td>
</tr>
<tr>
<td></td>
<td>Not OA</td>
<td>0.670</td>
<td>0.703</td>
<td>0.741</td>
<td>0.699</td>
</tr>
</tbody>
</table>

Discussion and Conclusions

Based on a sample of 10,041 journals, this study presents a comparison between Altmetrics metrics and citation metrics to discuss the applicability of Altmetrics in journal evaluation.

First on the overall applicability of Altmetrics metrics, according to our findings, the Altmetrics metrics calculated with the Altmetrics data of articles can reflect both of the short-term and long-term influence of journals to some extent, and it does a better job in the former one. In this study we find that Readers Sum, Accounts Sum and Altmetrics Scores can get evaluation results highly correlated to traditional results. Also, these metrics have good discrimination on the academic quality of journals. Because the influence of journal in the Internet is more complicated, Altmetrics are more able to reflect the influence which are exposed in the Internet. It can indicate the influence in the Internet in addition to reflecting the academic influence. Therefore, Altmetrics metrics are helpful to point out journals with greater comprehensive impact in a short period.

Second on Altmetrics data sources, the choice of data sources will affect the journal influence it reflects. Altmetrics data mainly comes from reference manager tools and social network. Data from reference manager tools can provide results more correlated to traditional results, while using data from social network is opposite. Further, data sources with a mature social network and fast update speed tend to be more able to indicate academic impact. One possible reason is that they have network structures, which are similar to citation data with citation network.

Third on the use of Altmetrics metrics, sum metrics can get results closer to traditional results compared with mean metrics. In most cases, sum metrics of Altmetrics metrics are more stable and are more able to reflect the academic influence of journals.

Fourth on the applicable scope of Altmetrics, currently it’s limited and not all journals are suitable to be evaluated with Altmetrics. In this study we find that Altmetrics is not applicable for specific disciplines or publishing areas. Altmetrics achieves better results in journals in Social Science and Health Science, instead of Natural Science especially journals in Computer Science and Mathematics. In addition, the Open Access status of journals will affect the evaluation results obtained by Altmetrics.

As for the future study on Altmetrics, one question worthy of attention is the difference between article level and journal level, since journals are the aggregation of articles while they show obvious difference in Altmetrics. Besides, the connotation and algorithm of Altmetrics should be emphasized. Since the usage of academic literatures on the Internet is very complex, the motives behind different behaviours can help to find out the characteristics and influence of journals. Because the data source of Altmetrics has a great influence on the evaluation, a scientific design of metrics for Altmetrics will be meaningful to use data from different sources effectively and obtain feasible method for the future scientific evaluation.
Acknowledgments
This work was supported by the National Social Science Fund of China [Grant No. 14BTQ067]. The authors would like to thank the anonymous reviewers for their detailed comments and valuable suggestions which helped to improve the paper.

References
Does Monetary Support Increase the Number of Scientific Papers?
An Interrupted Time Series Analysis

Yaşar Tonta

yasartonta@gmail.com
Hacettepe University, Ankara (Turkey)

Abstract
One of the main indicators of scientific production is the number of papers published in scholarly journals. Turkey ranks 18th place in the world based on the number of scholarly publications. The objective of this paper is to find out if the monetary support program initiated in 1993 by the Turkish Scientific and Technological Research Council (TÜBİTAK) to incentivize researchers and increase the number, impact and quality of international publications has been effective in doing so. We analyzed some 390,000 publications with Turkish affiliations listed in the Web of Science (WoS) database between 1976 and 2015 along with about 157,000 supported ones between 1997 and 2015. We used the interrupted time series analysis technique to test if TÜBİTAK’s support program helped increase the number of publications. It appears that the support program has had negligible impact on the increase of the number of papers. We discuss the implications of findings along with the possible causes of the ineffectiveness of the support program.

Conference Topic
Country level studies

Introduction
The number of scholarly papers and citations thereto are indirect indicators of the level of scientific development of countries. The number of scholarly papers with Turkish affiliations listed in citation indexes has increased tremendously over the years and Turkey ranks 18th in the world in terms of number of publications. Over 36,000 papers were published in 2015 alone, although their scientific impact in terms of the number of citations they gather is well below the average of the world, the European Union (EU) and the OECD countries.

In 1993, the Turkish Scientific and Technological Research Council (TÜBİTAK) has initiated a monetary support program (UBYT) to incentivize researchers and increase the number, impact and quality of international publications authored by Turkish researchers. Considerable percentages of papers with Turkish affiliations were supported in the early years of this program, even though the rate of support has gradually decreased (to c. 30%) over the years due to the steep increase in the number of published papers with Turkish affiliations.

As part of the program, some 157,000 publications (93% of which were papers/articles) were supported between 1997 and 2015. The amount of support paid for each paper has been determined on the basis of the impact factor of the journal in which it was published.

The total amount of support was about 124 million Turkish Liras (in 2015 current prices; equal to c. 35 million USD). The number of papers supported, the total number of publications, and the amount of support increased four-, 10- and 13-fold, respectively, during this period.

The support program has been in place for almost a quarter century. Yet, its impact has not been evaluated in the past. We have been asked by TÜBİTAK to evaluate the effectiveness of the program and given the payment records of 157,000 supported publications. They included, among others, journal information (name, year, its class based on Journal Citation Reports’ subject categories), type of contribution (e.g., article, review) and the amount of support.

Based on the payment records provided, the characteristics (i.e., impact factors) of journals in which supported papers with Turkish affiliations appeared have been analyzed, the functioning of the support algorithm has been studied, and the effectiveness of the overall
support program has been evaluated. Findings indicate that the authors of mediocre papers published in journals with relatively low impact factors have mostly been supported due to the use of skewed distributions of journal impact factors in determining the amount of support. The existing support algorithm, on the other hand, does not seem to function as conceived. This paper presents only the findings of the interrupted time series analysis with a view to find out if the support program has had any impact on the increase of the number of papers with Turkish affiliations. It is organized as follows: The Literature Review section briefly discusses the findings of relevant studies including those that provide some background on the Turkish case. The Data and Method section describes the data used and provides information on interrupted time series analysis. The detailed findings are presented thereafter (Findings and Discussion), followed by Conclusions and Recommendations.

Literature Review
Performance-based research funding systems (PRFSs) came into being in 1980s. Based on rewarding the outputs, the rationale of PRFSs is to provide more support to institutions (or individuals) with higher performances so that the ones with lower performances will strive to improve theirs in order to get more support (Herbst, 2007, p. 90). Yet, it is not clear if PRFSs based on outputs and competition increase the scientific productivity and the impact of outputs. In a relatively recent study comparing PRFSs and outputs of eight countries, countries with less competitive PRFSs such as Denmark turned out to be as effective as the ones with more competitive PRFSs such as the UK and Australia (Auranen & Nieminen, 2010, p. 830). Some researchers drew attention to the potential “side effects” of PRFSs based on competition, as they tend to “homogenize” research outputs, discourage experiments using new approaches, and reward researchers playing “safe” even though their contributions may not have any societal impact (Geuna & Martin, 2003, p. 296). The idea of increasing productivity on the basis of outputs and competition seems more complicated than what decision-makers have initially thought (Auranen & Nieminen, 2010).

There are mainly two types of PRFSs in use: (1) the ones based on peer review or informed peer review supported with bibliometric measures; and (2) the ones based solely on bibliometric measures such as journal impact factors. The UK’s Research Excellence Framework (REF) is the largest research assessment system in the world (De Boer et al., 2015, p. 113). Based on peer review, REF has been used since 1986 to distribute funds to research institutes and universities on the basis of their performances. Despite their shortcomings, PRFSs based on bibliometric measures only are on the rise, as they are, in comparison to peer review, easier and less costly to apply as a “proxy” to assess performance. Therefore, they tend to get preferred by increasingly more countries lately.

PRFSs and publication support systems based on bibliometric measures generally use the number of papers published in refereed journals and their impact in terms of citations as the main criteria to determine the research institutes and researchers to be supported. Impact factors (IF) and article influence scores (AIS) of journals are the two most commonly used metrics. Journal IF was originally proposed by the late Eugene Garfield (1972) to help librarians in their selection of journals for subscription. It is an indicator of the quality of a journal in general and measures the citation impact of an “average” paper published therein. It does not say anything about the quality of an individual paper in that journal and how many citations, if any, it would gather in a certain period of time after its publication (e.g., two years). Citation distributions used to calculate the IFs of journals are quite skewed, indicating that few papers published in a given journal get cited much more frequently while the majority get unnoticed or rarely cited (Marx & Bornmann, 2013). This is the case even for the most prestigious journals with the highest IFs such as *Nature* (IF=38) and *Science* (IF=35). As high as 75% of articles published in these journals get cited fewer times than their journal IFs.
indicate (Larivière et al., 2016, p. 4, Table 2). Journal IFs vary by scientific discipline, too, as the number of researchers in each field, publication types (i.e., journal articles as opposed to books) and scholarly communication patterns tend to differ. In general, some 9%-10% of all the articles listed in Web of Science collect 44% of the total number of citations (Albarrán, Crespo, Ortuño & Ruiz-Castillo, 2011). More importantly, there exists no positive relationship between the number of citations an article gets and the IF of the journal in which it is published (Zhang, Rousseau & Sivertsen, 2017, p. 14), and a large body of literature detailing the shortcomings of the use of journal IFs as a performance measure is readily available (e.g., Seglen, 1997; Glänzel & Moed, 2002; Van Raan, 2005; Marx & Bornmann, 2013; Casadevall & Fang, 2012; Wouters et al., 2015). Yet, rather than checking the number of citations to the papers of individual researchers, PRFSs based on bibliometric measures continue to use journal IFs to assess the performance of individuals. Journal IFs are quite misleading in predicting the number of citations that any given article might get. What follows are a few examples of PRFS using journal IFs as a research assessment tool.

PRFSs are reviewed by several researchers (e.g., Geuna & Martin, 2003; European Commission, 2010; Hicks, 2012; De Boer et al., 2015). Most EU countries, Norway, USA, Australia, New Zealand and China have some PRFSs in place. We provide a few examples of PRFSs that either solely use journal IF or use it in combination with peer review (excluding the ones based only on peer review such as REF in the UK).

Italy uses a PRFS where an expert panel decides whether to use citation analysis or peer review (or both) for each publication. Universities are ranked on the basis of a quality score consisting of citations and other journal metrics, which determine the amount of support each university gets. Some 30% of the research funds are distributed according to the outcome of this evaluation (Abramo, D’Angelo & Di Costa, 2011, p. 930; Abramo & D’Angelo, 2016, p. 2055; Abramo & D’Angelo, 2011, p. 348).

Similarly, Spain uses a mixed system, although researchers are encouraged to publish in journals that are listed in the top quarters of JCR’s subject categories. Researchers who publish in such journals receive monetary support that ranges somewhere between 3% and 15% of their monthly salaries (Osuna, Cruz-Castro & Sanz-Menéndez, 2011).

A number of countries such as Czech Republic, China, Finland, and Australia use journal IF exclusively to support research institutes and individual researchers. Finland, for instance, linked journal IF directly with research support by legislation (Adam, 2002, p. 727). Similarly, Australia and the Czech Republic make direct linkage between research evaluation and funding by counting scholarly outputs and assigning a score to each on the basis of bibliometric measures. These scores are then used to determine the amount of monetary support and papers that appear in refereed journals or journals with relatively higher IFs get the highest scores (Butler, 2004; Butler, 2003, p. 147-151; Good et al. 2015, p. 92, 96, Table 3). Norway also has a similar system based on weighting journals on the basis of various criteria and created three different journal lists (Schneider, 2009). China, on the other hand, uses journal IF most comprehensively in that academic recruitments and promotions, university rankings (and the amount of research support they get), support of Chinese journals that are listed in Chinese Citation Indexes all rely on journal IFs. The procedure seems to have been automated, as a researcher publishing in a journal with a certain IF knows how much support s/he would get. For instance, the author of a paper published in a journal with IF higher than 15 receives 300,000 Yuan (c. 43,000 USD) (Shao & Shen, 2012)!

Turkey is no exception: journal IFs are considered as an indicator of quality and have been used as an important criterion in academic promotions since early 1990s. In addition to individual universities, TÜBİTAK has initiated a nationwide monetary support system based exclusively on journal IFs. Journals classified under Q1, Q2, etc. in JCR’s subject categories have been used to determine the monetary compensation. More recently (2016), Turkish
Higher Education Council (HEC) started a new support scheme based mostly on journal IFs and the faculty whose scores are above a certain threshold in terms of number of academic activities (mostly publications) during the previous year get an additional 10% to 15% on top of their regular monthly salaries throughout the year.

It should be noted that performance-based research funding and publication support systems based on quantitative measures tend to have some adverse effects. Researchers seem to adjust to the requirements very easily and change their publication patterns and behaviors. Such systems are prone to "gaming", too, and researchers become more "opportunistic" (e.g., publication "inflation") and less ethical (e.g., "fake" citations) in time. Unintended consequences of PRFSs in several countries (e.g., Australia, Czech Republic, and Spain) were reported in the literature (Butler, 2003; Butler, 2004; Good et al., 2015; Osuna, Cruz-Castro & Sanz-Menéndez, 2011). For example, more papers tend to get published in journals with relatively lower IFs. A similar trend has also been observed in Turkey (Yurtsever et al., 2001, 2002; Önder et al., 2008; Kamalski et al., 2017, p. 298-301). As the Goodhart’s Law states, “When a measure becomes a target, it ceases to be a good measure”.1

It should also be noted that correlation between competitive PRFSs and the research productivity is not clear-cut (Auranen & Nieminen, 2010, p. 831). Excessive competition seems to reduce the time and energy otherwise to be expended for research. In this paper, we test the conjecture if TÜBİTAK’s publication support system has had an impact on the increase of number of publications listed in citation indexes with Turkish affiliations.

Data Sources and Method

We performed a search on Web of Science (WoS) (December 19, 2016) to identify all the publications with Turkish affiliations listed in Science Citation Index (SCI), Social Sciences Citation Index (SSCI) and Arts & Humanities Citation Index (A&HCI) between 1976 and 2015. More than 390,000 records were retrieved, 81% of which were full papers (articles) while the rest were other types of publications (e.g., reviews, notes, and letters to the editor).

TÜBİTAK provided the payment data for about 157,000 supported publications (93% of which were papers). These records were first cleaned, then coded as either “full papers” (articles) or “other” types of publications, classified under various criteria (e.g., year, class of journal, amount of support paid), ranked and combined, if necessary.

We used MS Excel and SPSS 23 for the detailed analysis of data and prepared both WoS and TÜBİTAK records for interrupted time series analysis outlined below (Interrupted, 2013).2

The interrupted time series (ITS) analysis technique (also known as quasi-experimental time series analysis or intervention analysis) is used in this paper to measure the impact of TÜBİTAK’s support program. ITS analysis measures if an “event” occurring at any given stage has an immediate or delayed effect on the time series data. For instance, an unexpected political development in a given country may increase the exchange rates, or a terrorist attack may reduce the number of tourists. These “events” (called “interventions”) may be planned or not planned. As ITS analysis is a quasi-experimental method, it is possible (by means of using a control group) to verify if the change has occurred because of the intervention. ITS analysis is based on the following statistical model:

\[
Y_t = \beta_{pre} + \beta_{post} + e_t
\]

1 See https://en.wikipedia.org/ s.v. “Goodhart’s Law”.
2 Time series data prepared for interrupted time series analysis can be had from the author.
3 This percentage should ideally be 0 (zero) in order for it to function as a true control group. Yet, we think that time series data prepared for interrupted time series analysis can be had from the author.
where $Y_t$ represents the $t$'th observation in the time series, $\beta_{\text{pre}}$ and $\beta_{\text{post}}$ represent the levels of series before and after the intervention, respectively, and $e_t$ is the error related with $Y_t$. The null hypothesis

$$H_0 = \beta_{\text{pre}} - \beta_{\text{post}} = 0$$

states that there is no statistically significant difference between the levels of series before and after the intervention (i.e., it has no impact on dependent variable (McDowall et al., 1980, p. 12). It is assumed that the parameters in time series models stay the same before and after the intervention and that no other events that affect the parameters take place. ITS analysis can be applied to both static and dynamic (“ergodic”) time series. The ARIMA model is used for non-static series whose arithmetic means, variances and co-variances change as time passes. This model is expressed as ARIMA ($p, d, q$) where $p, d$ and $q$ represent the autoregressive operator (AR), the integrated operator (I), and the moving average operator (MA), respectively. If time series data is not stationary ($d$), it will first be made stationary to make its mean and variance constant over the years studied.

We have WoS data of publications with Turkish affiliations (1976-2015) and data of supported publications by TÜBİTAK (1997-2015). The program (“intervention”) started in 1993 and enough data points exist both before (1976-1992) and after (1993-2015) the intervention so as to be able to apply ITS analysis to time series data (Cochrane, 2002, p. 7-8). As relatively fewer researchers benefited from the support program in the early years, we thought that the effect of the program might be observed with some delay (lag). Therefore, we measured its impact one (1994), four (1997) and 10 years (2003) after of its start.

We have no data on papers (full articles) whose authors have not been supported. However, a relatively small group of authors of other types of contributions can function as a control group, as only 3% of the total amount of support on average was set aside for such contributions even though 19% of publications were of such nature. The authors of other types of contributions were paid half of what the authors of the full papers were, and a mere 1% of the support budget was allocated to them in 2013, for example. In other words, we can find out if TÜBİTAK’s support program has had any impact on the increase in the number of papers by comparing it with that of other types of contributions. If the number of other types of contributions that were not well supported did not increase but the number of papers supported increased, we can deduce that the source of the impact was the support program. Conversely, if, despite lack of support, the number of other types of contributions increased along with the number of papers receiving full monetary support, then the increase in the latter cannot be attributed to the program, suggesting that some factor(s) other than the support program may have played a role in this increase.

**Findings and Discussion**

The descriptive data about the number of papers and the total number of publications originating from Turkey are presented in Table 1 and Fig. 1. The rate of increase is quite steep, especially starting from 2000s. This rate of increase made Turkey in those years one of the fastest growing countries in the world in terms of number of papers, and Turkey moved up the ladder very quickly from 45th in 1983 to 25th in 1999 to 18th in 2008 in the world, contributing to 1.56% of the overall scientific production in the world.

A considerable percentage of these publications were supported by TÜBİTAK’s support program when it was first initiated in 1993. However, the support program seems to have not

---

3 This percentage should ideally be 0 (zero) in order for it to function as a true control group. Yet, we think that it can be used as a control group with some caution and the generalization should be interpreted accordingly.
kept up with the pace of increase of papers and the percentage of papers supported went down from 70% in early 2000s to below 30% in recent years (Table 2, Fig. 2).

### Table 1. Number of publications with Turkish affiliations (1976-2015)

<table>
<thead>
<tr>
<th>Year</th>
<th>Papers N</th>
<th>Papers %</th>
<th>Other N</th>
<th>Other %</th>
<th>Total N</th>
<th>Year</th>
<th>Papers N</th>
<th>Papers %</th>
<th>Other N</th>
<th>Other %</th>
<th>Total N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1976</td>
<td>216</td>
<td>80</td>
<td>53</td>
<td>20</td>
<td>269</td>
<td>1996</td>
<td>3359</td>
<td>84</td>
<td>623</td>
<td>16</td>
<td>3982</td>
</tr>
<tr>
<td>1977</td>
<td>229</td>
<td>72</td>
<td>91</td>
<td>28</td>
<td>320</td>
<td>1997</td>
<td>3844</td>
<td>83</td>
<td>796</td>
<td>17</td>
<td>4640</td>
</tr>
<tr>
<td>1978</td>
<td>272</td>
<td>72</td>
<td>108</td>
<td>28</td>
<td>380</td>
<td>1998</td>
<td>4460</td>
<td>82</td>
<td>1001</td>
<td>18</td>
<td>5461</td>
</tr>
<tr>
<td>1979</td>
<td>256</td>
<td>71</td>
<td>106</td>
<td>29</td>
<td>362</td>
<td>1999</td>
<td>5201</td>
<td>83</td>
<td>1078</td>
<td>17</td>
<td>6279</td>
</tr>
<tr>
<td>1980</td>
<td>343</td>
<td>74</td>
<td>123</td>
<td>26</td>
<td>466</td>
<td>2000</td>
<td>5462</td>
<td>84</td>
<td>1059</td>
<td>16</td>
<td>6521</td>
</tr>
<tr>
<td>1981</td>
<td>299</td>
<td>73</td>
<td>110</td>
<td>27</td>
<td>409</td>
<td>2001</td>
<td>6684</td>
<td>84</td>
<td>1271</td>
<td>16</td>
<td>7955</td>
</tr>
<tr>
<td>1982</td>
<td>315</td>
<td>70</td>
<td>132</td>
<td>30</td>
<td>447</td>
<td>2002</td>
<td>8985</td>
<td>86</td>
<td>1434</td>
<td>14</td>
<td>10419</td>
</tr>
<tr>
<td>1983</td>
<td>354</td>
<td>72</td>
<td>141</td>
<td>28</td>
<td>495</td>
<td>2003</td>
<td>10662</td>
<td>84</td>
<td>1978</td>
<td>16</td>
<td>12640</td>
</tr>
<tr>
<td>1984</td>
<td>420</td>
<td>77</td>
<td>129</td>
<td>23</td>
<td>549</td>
<td>2004</td>
<td>13199</td>
<td>84</td>
<td>2488</td>
<td>16</td>
<td>15687</td>
</tr>
<tr>
<td>1985</td>
<td>447</td>
<td>76</td>
<td>145</td>
<td>24</td>
<td>592</td>
<td>2005</td>
<td>14194</td>
<td>83</td>
<td>2877</td>
<td>17</td>
<td>17071</td>
</tr>
<tr>
<td>1986</td>
<td>506</td>
<td>77</td>
<td>151</td>
<td>23</td>
<td>657</td>
<td>2006</td>
<td>15070</td>
<td>79</td>
<td>4099</td>
<td>21</td>
<td>19169</td>
</tr>
<tr>
<td>1987</td>
<td>588</td>
<td>77</td>
<td>174</td>
<td>23</td>
<td>762</td>
<td>2007</td>
<td>17853</td>
<td>80</td>
<td>4414</td>
<td>20</td>
<td>22267</td>
</tr>
<tr>
<td>1988</td>
<td>672</td>
<td>75</td>
<td>227</td>
<td>25</td>
<td>899</td>
<td>2008</td>
<td>19327</td>
<td>82</td>
<td>4379</td>
<td>18</td>
<td>23706</td>
</tr>
<tr>
<td>1989</td>
<td>829</td>
<td>80</td>
<td>209</td>
<td>20</td>
<td>1038</td>
<td>2009</td>
<td>21655</td>
<td>82</td>
<td>4627</td>
<td>18</td>
<td>26282</td>
</tr>
<tr>
<td>1990</td>
<td>912</td>
<td>78</td>
<td>261</td>
<td>22</td>
<td>1173</td>
<td>2010</td>
<td>22833</td>
<td>83</td>
<td>4760</td>
<td>17</td>
<td>27593</td>
</tr>
<tr>
<td>1991</td>
<td>1134</td>
<td>80</td>
<td>290</td>
<td>20</td>
<td>1424</td>
<td>2011</td>
<td>23588</td>
<td>82</td>
<td>5325</td>
<td>18</td>
<td>28913</td>
</tr>
<tr>
<td>1992</td>
<td>1351</td>
<td>77</td>
<td>406</td>
<td>23</td>
<td>1757</td>
<td>2012</td>
<td>25254</td>
<td>82</td>
<td>5607</td>
<td>18</td>
<td>30861</td>
</tr>
<tr>
<td>1993</td>
<td>1519</td>
<td>76</td>
<td>482</td>
<td>24</td>
<td>2201</td>
<td>2013</td>
<td>26526</td>
<td>79</td>
<td>7200</td>
<td>21</td>
<td>33726</td>
</tr>
<tr>
<td>1994</td>
<td>1754</td>
<td>73</td>
<td>643</td>
<td>27</td>
<td>2297</td>
<td>2014</td>
<td>27242</td>
<td>79</td>
<td>7315</td>
<td>21</td>
<td>34557</td>
</tr>
<tr>
<td>1995</td>
<td>2233</td>
<td>72</td>
<td>885</td>
<td>28</td>
<td>3118</td>
<td>2015</td>
<td>28662</td>
<td>79</td>
<td>7530</td>
<td>21</td>
<td>36192</td>
</tr>
</tbody>
</table>

**Total / Avg.** 318709 81 74727 19 393436
Table 2. Number of papers supported by TÜBİTAK (1997-2015)

<table>
<thead>
<tr>
<th>Year</th>
<th># of papers supported by TÜBİTAK</th>
<th># of papers with Turkish affiliations (WoS)</th>
<th>Percentage supported (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>2247</td>
<td>3844</td>
<td>58</td>
</tr>
<tr>
<td>1998</td>
<td>2657</td>
<td>4460</td>
<td>60</td>
</tr>
<tr>
<td>1999</td>
<td>3088</td>
<td>5201</td>
<td>59</td>
</tr>
<tr>
<td>2000</td>
<td>3298</td>
<td>5462</td>
<td>60</td>
</tr>
<tr>
<td>2001</td>
<td>4216</td>
<td>6684</td>
<td>63</td>
</tr>
<tr>
<td>2002</td>
<td>5888</td>
<td>8985</td>
<td>66</td>
</tr>
<tr>
<td>2003</td>
<td>7517</td>
<td>10662</td>
<td>71</td>
</tr>
<tr>
<td>2004</td>
<td>9511</td>
<td>13199</td>
<td>72</td>
</tr>
<tr>
<td>2005</td>
<td>7036</td>
<td>14194</td>
<td>50</td>
</tr>
<tr>
<td>2006</td>
<td>8122</td>
<td>15070</td>
<td>54</td>
</tr>
<tr>
<td>2007</td>
<td>10551</td>
<td>17853</td>
<td>59</td>
</tr>
<tr>
<td>2008</td>
<td>10411</td>
<td>19327</td>
<td>54</td>
</tr>
<tr>
<td>2009</td>
<td>11554</td>
<td>21655</td>
<td>53</td>
</tr>
<tr>
<td>2010</td>
<td>11592</td>
<td>22833</td>
<td>51</td>
</tr>
<tr>
<td>2011</td>
<td>9574</td>
<td>23588</td>
<td>41</td>
</tr>
<tr>
<td>2012</td>
<td>10641</td>
<td>25254</td>
<td>42</td>
</tr>
<tr>
<td>2013</td>
<td>10203</td>
<td>26526</td>
<td>38</td>
</tr>
<tr>
<td>2014</td>
<td>10257</td>
<td>27242</td>
<td>38</td>
</tr>
<tr>
<td>2015</td>
<td>8014</td>
<td>28662</td>
<td>28</td>
</tr>
<tr>
<td>Total</td>
<td>146377</td>
<td>318709</td>
<td>46</td>
</tr>
</tbody>
</table>

Fig. 2. Number of papers listed in WoS with Turkish affiliations and supported by TÜBİTAK (1997-2015)
The detailed analysis of changes in TÜBİTAK’s support policies over the years is beyond the confines of this paper. Instead, we concentrate on whether TÜBİTAK’s support program has actually played a role in the steep rate of increase of papers by Turkish researchers. The time path of the number of papers listed in the Web of Science (WoS) originating from Turkey between 1976 and 2015 is given below (Fig. 3). The intervention point (1993) is marked on the graph. As there exists a trend of increase in the number of papers both before and after the intervention, we took the difference of the time series from the 1st level \((d=1)\) to make it stationary. Consequently, the auto-correlation function (ACF) and partial auto-correlation function (PACF) of the time series became static within the confidence intervals (Fig. 4).

Fig. 3. Time path of papers with Turkish affiliations (1976-2015)

Fig. 4. Correlograms of autocorrelation (ACF) and partial autocorrelations (PACF) functions
We then defined ARIMA (1,1,0) model for interrupted time series data and wanted to see the impact of TÜBİTAK’s support program in 1994, 1997 and 2003 (after one, four and 10 years of its start, respectively). The test statistic of the ARIMA model shows that the defined model is suitable for the time series data ($\chi^2 = 23.531$, DF = 17, $p = .133$) (Table 3). The parameters of the ARIMA model (estimates, SE, t- and p- values) are given in Table 4. The ARIMA Model did not produce statistically significant results (coefficient = .153, SE = .170, $t = 0.899$, $p = .375$). The coefficient for “Time series” in Table 4 gives the slope of the regression line before the intervention (14.051), which is used to analyze the different time points by taking into account the existing trend in data before calculating the effect of the intervention. The coefficient for “Before/after Support Program” represents the slope of y-axis when $x$ is equal to 0 (zero) and is used to measure the effect of the intervention in later time points. The coefficient for “Effect” (29.091) gives the difference between slopes before and after the intervention. By adding this difference to the value of pre-intervention slope (14.051), the value of the post-intervention slope (44.142) can be calculated (Interrupted, 2013).

<table>
<thead>
<tr>
<th>Table 3. Test statistic (Ljung Box)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>Makale sayısı-Model_1</td>
</tr>
</tbody>
</table>

Table 4. ARIMA Model Parameters

<table>
<thead>
<tr>
<th># of papers</th>
<th>Model 1</th>
<th># of papers</th>
<th>transformation</th>
<th>Constant</th>
<th>Lag</th>
<th>AR</th>
<th>1</th>
<th>.153</th>
<th>.170</th>
<th>.899</th>
<th>.375</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>No</td>
<td>No</td>
<td>0</td>
<td>14.051</td>
<td>29.910</td>
<td>.470</td>
<td>.642</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In order to see the effect of the support program on the number of papers with Turkish affiliations, we continued with this model. The slopes of pre- and post-intervention are the same for all analyses. It is possible to see the direct effect of the intervention on the number of papers with Turkish affiliations (Table 5). According to the model, an additional 554 papers were published in 1994 because of the support program. However, the effect of the support program is not statistically significant ($p = .157$). The delayed effect of the program has not been materialized in later years, either, as additional number of papers published due to the program were limited (651 papers in 1997, and 826 in 2003) and the effect is not statistically significant ($p > .05$). As the effect of the program has been negligible, the formula of the effect of the intervention is not given.

Despite the fact that other types of contributions have been supported very little during the period of analysis, the rate of their increase is greater than that of generously supported papers (see Fig. 1). As a control group, the rate of continuous increase in other types of publications seems to confirm the results of the interrupted time series analysis. For instance, some 4,000-
7,000 other types of publications have been published annually in recent years, of which only a few hundreds got supported. Yet, the number of other publications continues to increase regardless of support, suggesting that TÜBİTAK’s support program is probably not the main factor causing the increase in the number of papers with Turkish affiliations.

Table 5. Values showing the delayed effect of TÜBİTAK’s support program

<table>
<thead>
<tr>
<th>Year</th>
<th>Predicted increase</th>
<th>SE</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1994</td>
<td>563.633</td>
<td>390.084</td>
<td>1.446</td>
<td>.157</td>
</tr>
<tr>
<td>1997</td>
<td>651.241</td>
<td>431.129</td>
<td>1.510</td>
<td>.140</td>
</tr>
<tr>
<td>2003</td>
<td>825.784</td>
<td>571.279</td>
<td>1.446</td>
<td>.157</td>
</tr>
<tr>
<td>2015</td>
<td>1,174.941</td>
<td>947.761</td>
<td>1.240</td>
<td>.224</td>
</tr>
</tbody>
</table>

It should be noted though that interrupted time series analysis has some limitations. The assumption that no other “event” or “events” occurred during the period of analysis that might have affected the time series data is one of them. For example, the prerequisite of having papers published in journals listed in citation indexes for academic promotion may have triggered this increase, as more than 90% of research in Turkey has been carried out in universities, and the number of academic personnel in universities has increased tremendously over the years. Moreover, in addition to the number of research personnel in universities, the number of papers may be increasing due to a number of other factors such as the number of researchers per 10,000 capita, and the share of R&D expenditures within the Gross National Product (GDP). As indicated earlier, even though some positive correlation between PRFSs and the number of papers has been observed, this may not necessarily point to a strong causality between the two. As was the case in Spain (Osuna, Cruz-Castro & Sanz-Menéndez, 2011), the number of papers with Turkish affiliations continues to increase perhaps not because of TÜBİTAK’s support program but because of other factors such as the growth in and the maturity of universities’ research systems including academic personnel.

We should also note that we carried out a multiple regression analysis and observed fairly strong correlation between the number of papers with Turkish affiliations and the number of supported papers. However, we decided not to report the results of the multiple regression analysis, as the Durbin-Watson statistic was rather small (0.921), probably indicating the existence of serial autocorrelation between variables and thereby making the results less reliable. This can to some extent be observed from Fig. 2: the correlation between number of papers with Turkish affiliations and the supported papers was positive and statistically significant between 1997 and 2006 whereas it was negative and not statistically significant between 2007 and 2015.

Conclusion

As part of TÜBİTAK’s support program, the authors of over 157,000 publications received more than 124 million Turkish Liras (in 2015 current prices, c. 35 million USD) as monetary support between 1997 and 2015. Yet, two thirds of all payments were less than 826 liras (or c. 230 USD). These “micropayments” might be one of the reasons why, according to the test results of the interrupted time series analysis, the program did not seem to have direct impact on the increase of the number of papers published by Turkish authors. It is likely that small amounts of payments were not much of an incentive for authors to publish more.

We should point out that the objective of the support program is not to increase the number of papers per se but to increase their impact and quality, as stated in the By-Law of TÜBİTAK’s support program (TÜBİTAK, 2016). Some authors may find the small payments satisfactory. Yet, if such small payments do not help achieve the program’s objectives, precautions should be taken to correct it. The support program seems to have functioned as a mechanism to transfer
small amounts of payments to authors without any considerable improvement in the impact and quality of the papers. Transaction costs of such small payments should be borne in mind as well as the costs of missed opportunities of increasing the impact and quality of papers.

Acknowledgments
This paper is based on a report (in Turkish) on the evaluation of TÜBİTAK’s Support Program of International Scientific Publications (UBYT), which is to be published by TÜBİTAK as a monograph (Tonta, in press). Research was carried out during my sabbatical year (2016-2017) at the School of Library and Information Science of Humboldt University in Berlin. I gratefully acknowledge the continuous support that I received throughout my stay from my family and colleagues who assumed some of my responsibilities as well as from my home and host institutions (Hacettepe University and Humboldt University) and TÜBİTAK. I thank all the people who provided the payments data; offered a pleasant work environment with access to information resources; and helped me in completing the research and writing the report.

References


Research competitiveness analysis refers to the measurement, comparison and analysis of the research status (i.e., strength and/or weakness) of different scientific research bodies (e.g., institutions, researchers, etc.) on different research fields. Improving research competitiveness analysis method can be conducive to accurately obtaining the research status of research fields and research bodies. This paper presents a method of evaluating the competitiveness of research institutions based on research topic distribution. The method uses the LDA topic model to obtain paper-topic distribution matrix to objectively assign the academic impact of papers (such as times of citation) to research topics. Then the method calculates the competitiveness of each research institution on each research topic with the help of institution-paper matrix. Finally, the competitiveness and the research strength and/or weakness of the institutions are defined and characterized. Case study shows that the method can lead to an objective and effective evaluation of the research competitiveness of given research institutions on given research field.

Conference Topic
The theory, method and principle of five metrics science concepts, that is, Bibliometrics, Informetrics, Scientometrics, Webometrics and Knowledgometrics.

Introduction
Research competitiveness analysis refers to the measurement, comparison and analysis of the research status of different scientific research bodies (e.g., institutions, researchers, etc.) on different research fields (Zhang, 2014). Improving competitiveness analysis methods can be conducive to obtaining the research status of research bodies, clarifying their strengths and/or weaknesses, and in turn promoting collaborative innovation among different research bodies and different research fields.

Generally, research competitiveness analysis is carried out based on research papers and involves three steps, i.e., research field (topic) recognition, competitiveness evaluation and competitiveness analysis (Gei, 2013). Research field recognition is important since different research fields are always not comparable, while topic recognition can result in a fine granular evaluation for strength and/or weakness characterization. For field recognition, a paper’s research field is usually determined based on partition standard provided by the scientific literature database providers, e.g., ESI, Incites, Wos, etc. (Chen & Shi, 2013; Dong, 2014; Li 2012; Cova, 2013). For research topic, it is often represented by keywords with high frequencies and their frequent combinations derived using certain analysis tools (e.g., CiteSpace) (Chen, 2006; Chen & Hu, 2012). When the fields (topics) are determined, the scientometrics criterions of papers are used for competitiveness evaluation of the corresponding institutions on the corresponding research fields (topics). Those criterions may include paper count, paper IFs, paper citation counts, etc. (Mkhnacheva, 2011; Morris, 2003; Small, 2009; Shibata, 2008). Finally, the ranking of competitiveness is given and the strengths and/or weaknesses are analysed (Liu, 2015; Small, 2014).

Those methods have made concrete progress on competitiveness evaluation; however, the following problems should be further considered. First, the mapping from papers to research fields and topics are too straightforward for a precise evaluation, since many multi-discipline papers cannot be simply partitioned into a unique research field, and a small set of frequent

* Corresponding Author
keywords may not be capable to represent a research topic. Second, in the current works, papers are all bounded to a unique research field, and papers relating to multiple research topics are considered contributing equally to each of the topics. However, this may not always intuitive since a paper always has certain main research points corresponding to one or more (but not all the relating) field(s) and/or topic(s).

Focusing on the mentioned problems, this paper presents a method of evaluating the competitiveness of research institutions based on research topic distribution. The method uses the LDA topic model to obtain paper-topic distribution matrix to objectively assign the academic impact of papers (such as the number of cited times) to research topics. Then the method calculates the impact of each research institution on each research topic with the help of institution-paper matrix. Further, the competitiveness scores of institutions are calculated and the research strength and/or weakness of the institutions are defined and derived. Finally, case study is carried out to show effectiveness of the proposed evaluation method. It is to be noted that in the proposed method, there is no need to distinguish research field and topic, since a research field can be viewed as a higher abstraction of research topics. That means by setting proper parameters, LDA can be used to model paper-field distribution.

**Evaluation Method**

The proposed method goes through the following steps for evaluation: 1) topic recognition, 2) impact allocation, 3) competitiveness measurement. We explained each step as follows.

**Topic recognition**

LDA is a document topic generation model (Blei, 2003). The model presumes that the words in the topic and the topics of the document are both subject to certain polynomial distributions. Hence generating a document can be seen as a repeated process of selecting a topic with a certain probability and then selecting a word in the topic with a certain probability. The model can be formally represented as $\Omega = \Phi \times \Theta$, where $\Omega$, $\Phi$ and $\Theta$ is document-word distribution, topic-word distribution and document-topic distribution respectively, $\times$ represents matrix multiplication, as demonstrated in Fig. 1. In Fig 1, we have 3 papers; each is composed of 2 topics and 3 words. The LDA model can be used to determine $\Theta$ for a set of documents by setting a proper topic number $n$.

\[
\begin{pmatrix}
\omega_{11} & \omega_{12} & \omega_{13} \\
\omega_{21} & \omega_{22} & \omega_{23} \\
\omega_{31} & \omega_{32} & \omega_{33}
\end{pmatrix} = \begin{pmatrix}
\varphi_{11} & \varphi_{12} \\
\varphi_{21} & \varphi_{22} \\
\varphi_{31} & \varphi_{32}
\end{pmatrix} \times \begin{pmatrix}
\theta_{11} & \theta_{12} & \theta_{13} \\
\theta_{21} & \theta_{22} & \theta_{23}
\end{pmatrix}
\]

**Figure 1 LDA Model of 3 papers, 2 topics and 3 words**

**Impact allocation**

Once we got the paper-topic distribution matrix $\Theta$, we can then allocate the academic impact of the papers to the relating topics using values in $\Theta$ as weights to get paper-impact matrix $\Gamma$ (that characterizes the impact of papers on topics). More formally, given $m$ papers, $n$ topics, paper-topic distribution matrix $\Theta = \{ \theta_{ij} | \theta_{ij} \in [0,1], i \in [1,m], j \in [1,n], \sum_{j} \theta_{ij} = 1 \}$, paper-impact vector $\mathbf{I} = [t_1, t_2, \ldots, t_m]^T$, the paper-impact matrix $\Gamma$ can be calculated as $\Gamma = \Theta^T \otimes \mathbf{I} = \{ \gamma_{ij} | \gamma_{ij} = \theta_{ij} \times t_j, \theta_{ij} \in \Theta^T, t_j \in \mathbf{I} \}$, where $\theta_{ij}$ is the weight of paper $i$ on topic $j$, $t_j$ is the impact indicator of paper $i$, $\gamma_{ij}$ is the calculated impact value of paper $i$ on topic $j$. Fig. 2 (a) shows an example with 3 papers and 2 topics.
In practice, a paper’s academic impact may relate to many aspects, e.g., times of citation, impact factor, and etc. We use a weighted composition of the aspects to make an overall evaluation of paper impact. That is, suppose we have \( l \) factors (that influence academic impact of a paper) whose values are given in the paper-impact matrix \( \Lambda = \{ \lambda_{ij} \mid i \in [1,n], j \in [1,l] \} \), the paper-impact vector is calculated as \( \mathbf{I} = \Lambda \times \mathbf{A} \), where \( \lambda_{ij} \) is the normalized impact value of paper \( i \) on factor \( j \). A \( \Lambda = \begin{bmatrix} 0.26 & 0.26 \\ 0.252 & 0.588 \\ 0.4 & 0.1 \end{bmatrix} \) \( \times \begin{bmatrix} 0.5 & 0.5 \\ 0.3 & 0.7 \\ 0.8 & 0.2 \end{bmatrix} \otimes \begin{bmatrix} 0.52 \\ 0.84 \\ 0.50 \end{bmatrix} \times \begin{bmatrix} 0.2 & 1 \\ 1 & 0.6 \end{bmatrix} \times \begin{bmatrix} 0.6 \\ 0.4 \end{bmatrix} \)

\[ \begin{bmatrix} 0.52 \\ 0.84 \end{bmatrix} = \begin{bmatrix} 0.7 & 0.2 \end{bmatrix} \]

\( \mathbf{I} \), the institution-impact matrix and \( \mathbf{A} \), the institution-impact vector.

Now we can measure the competitiveness of the relating institutions based on institution-impact matrix \( \Xi \). First, we calculate the average impact of all the institutions on each topic: \( \xi = \langle \xi_1, \xi_2, \ldots, \xi_n \rangle \), where \( \forall j \in [1,n], \xi_j = \frac{1}{v} \sum_{i=1}^{v} \xi_{ij} \), where \( \xi_{ij} \in \Xi \). Then, we calculate the difference matrix \( \Delta \) between \( \Xi \) and \( \xi \) as \( \Delta = \Xi \ominus \xi = \{ \delta_{ij} \mid \delta_{ij} = \xi_{ij} - \xi_j, \xi_{ij}, \xi_j \in \Xi, \xi_j \in \xi \} \), as exemplified in Fig. 4.

\[ \Delta = \begin{bmatrix} 0.18 & -0.02 \\ 0.04 & 0.14 \\ -0.22 & -0.12 \end{bmatrix} \begin{bmatrix} 0.652 & 0.688 \\ 0.512 & 0.848 \\ 0.252 & 0.588 \end{bmatrix} \ominus \begin{bmatrix} 0.472 & 0.708 \end{bmatrix} \]

Figure 2 Example of (a) calculation of paper-impact matrix and (b) calculation of paper-impact vector

Figure 3 Calculation of institution-impact matrix

Competitiveness measurement

The difference matrix \( \Delta \) can then be used to assess the research status of the institutions. Given a threshold \( \tau \) and a percentage \( p \), we can define, for example, 1) if \( \delta_{ij} \geq \tau \), then
research topic \( j \) is the strength of institution \( i \); 2) if \( \delta_{ij} \leq -\tau \), then research topic \( j \) is the weakness of institution \( i \); for a certain topic \( j \), after ranking the institutions according to \( \delta_{ij} \), let \( R_{ij} \) denote the ranking of institution \( i \) on topic \( j \). 3) if \( R_{ij} \leq \rho \), then institution \( i \) is leading the research on topic \( j \), and many other rules. Based on the definition, as to our example in Fig. 4, if we set \( \tau = 0.1 \), then we can say that topic 1 and 2 is the strength of institution 1 and 2 respectively; both topic 1 and 2 are the weaknesses of institution 3.

**Case Study**

In this section, we use case study to verify the effectiveness of the proposed evaluation method from two perspectives: 1) the rationality of LDA model, 2) the effectiveness of competitiveness evaluation.

For LDA model, we want to see whether various research topics can be identified by the model. For the purpose, we used \( SU = \) "computer science" and \( TS = \) "operating system"; \( SU = \) "computer science" and \( TS = \) "information security"; \( SU = \) "computer science" and \( TS = \) "artificial intelligence*"; \( SU = \) "computer science" and \( TS = \) "computer graphics*"; \( SU = \) "computer science" and \( TS = \) "software engineer*" as search strategies to search and download bibliographic data from Science Citation Index Expanded (SCI-EXPANDED) database. The data relates to 28421 research papers in five research topics of computer science, i.e., operating system (OS), information security (IS), artificial intelligence (AI), computer graphics (CG) and software engineering (SE). Then we extracted keywords from the data and input them to the LDA algorithm. By setting the topic number as 5, we got a paper-topic matrix of 28421 rows and 5 columns.

Fig. 5 demonstrates the topic distribution of a few of papers (21 papers) as an example. It can be seen that all the papers to some extent relates to all the topics. To verify the LDA model’s ability of identifying topics, we assigned the LDA topic with largest probability as a paper’s research topic (e.g., the topic of \( P_1 \) is \( T_1 \)) and recorded all the correspondences between the papers (in various research topics) and the LDA topics. The results illustrated in Table 1 show the correspondences of AI to \( T_4 \), CG to \( T_0 \), IS to \( T_2 \), OS to \( T_1 \) and SE to \( T_3 \). Table 2 shows the 5 most representative keywords of each topic. From the table we can see that the topics can be easily interpreted by a field expert, and the interpretation in the table can perfectly support the correspondences in Table 1.

<table>
<thead>
<tr>
<th>( T_1 )</th>
<th>( T_2 )</th>
<th>( T_3 )</th>
<th>( T_4 )</th>
<th>( T_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>0.459833</td>
<td>0.1641149359</td>
<td>0.163627393</td>
<td>0.075640847</td>
</tr>
<tr>
<td>F2</td>
<td>0.2109333399</td>
<td>0.1118833359</td>
<td>0.10178201</td>
<td>0.110413023</td>
</tr>
<tr>
<td>F3</td>
<td>0.1043950198</td>
<td>0.388316175</td>
<td>0.138986738</td>
<td>0.018975894</td>
</tr>
<tr>
<td>F4</td>
<td>0.110982605</td>
<td>0.157618074</td>
<td>0.123854964</td>
<td>0.137053852</td>
</tr>
<tr>
<td>F5</td>
<td>0.29472698</td>
<td>0.178078219</td>
<td>0.121781835</td>
<td>0.12775185</td>
</tr>
<tr>
<td>F6</td>
<td>0.18228132</td>
<td>0.140132938</td>
<td>0.16144473</td>
<td>0.407280753</td>
</tr>
<tr>
<td>F7</td>
<td>0.144154499</td>
<td>0.15220457</td>
<td>0.162594321</td>
<td>0.20749411</td>
</tr>
<tr>
<td>F8</td>
<td>0.1415475</td>
<td>0.15218519</td>
<td>0.162550589</td>
<td>0.20742155</td>
</tr>
<tr>
<td>F9</td>
<td>0.186451932</td>
<td>0.13120632</td>
<td>0.179222426</td>
<td>0.304441824</td>
</tr>
<tr>
<td>F10</td>
<td>0.23771837</td>
<td>0.136623345</td>
<td>0.170003951</td>
<td>0.229936012</td>
</tr>
<tr>
<td>F11</td>
<td>0.15300924</td>
<td>0.13925322</td>
<td>0.172329249</td>
<td>0.254342621</td>
</tr>
<tr>
<td>F12</td>
<td>0.14117715</td>
<td>0.13575247</td>
<td>0.176797488</td>
<td>0.270478613</td>
</tr>
<tr>
<td>F13</td>
<td>0.26396793</td>
<td>0.176706821</td>
<td>0.263878737</td>
<td>0.173821759</td>
</tr>
<tr>
<td>F14</td>
<td>0.19095304</td>
<td>0.151840357</td>
<td>0.206534308</td>
<td>0.28398887</td>
</tr>
<tr>
<td>F15</td>
<td>0.15966551</td>
<td>0.13728359</td>
<td>0.275413981</td>
<td>0.20483017</td>
</tr>
<tr>
<td>F16</td>
<td>0.19041626</td>
<td>0.16297473</td>
<td>0.198218203</td>
<td>0.28918184</td>
</tr>
<tr>
<td>F17</td>
<td>0.139412831</td>
<td>0.162735655</td>
<td>0.198217758</td>
<td>0.289196977</td>
</tr>
<tr>
<td>F18</td>
<td>0.18414759</td>
<td>0.162973784</td>
<td>0.198219931</td>
<td>0.289196514</td>
</tr>
<tr>
<td>F19</td>
<td>0.122749078</td>
<td>0.166316765</td>
<td>0.195497961</td>
<td>0.69627068</td>
</tr>
<tr>
<td>F20</td>
<td>0.15989825</td>
<td>0.098189067</td>
<td>0.264627364</td>
<td>0.140891816</td>
</tr>
<tr>
<td>F21</td>
<td>0.102269778</td>
<td>0.0995542631</td>
<td>0.494964601</td>
<td>0.160605681</td>
</tr>
</tbody>
</table>

*Figure 5 An example of paper-topic matrix*
For the effectiveness of competitiveness evaluation, we used AI ranking as example and compared the ranking resulted from our method with two other rankings. AI was chosen since there are numbers of public available rankings released by various third parties that can be compared. We reviewed three rankings, i.e., US News ranking 2016, CS Ranking 2016, and WUZHEN Ranking 2016. We found that the three rankings were similar with each other. For example, the Top 3 institutions in all the 3 rankings are Stanford University, Carnegie Mellon University and MIT. The Top 10 institutions in the 3 ranking also have many overlapping: in total only 16 institutions make up the 3 Top 10 lists. That is, in the 30 institutions of the 3 Top 10 lists, more than 40% of the institutions are overlapped. Consequently, we chose the most widely acknowledged ranking, i.e., US News ranking 2016, as our first reference for comparison. For the other reference, we want to compare the topic based method with the traditional method. That is, after topic recognition, we removed the concept of multi-disciplines from the paper-topic matrix \( \Theta \) so that each paper only contributes to the topic with largest probability. More formally, \( \forall \theta_j \in \Theta, \theta_j = 1, \text{if } \forall j \neq k, \theta_j > \theta_k ; \theta_j = 0, \text{else.} \) For example, the distribution of \( P1 \) on \( T1 \) to \( T5 \) in Fig. 5 will be 1, 0, 0, 0, 0. The following steps were all the same with the proposed method and the results were acquired for comparison. During the case study, citation counts and impact factors of papers were used for paper-impact vector calculation, and weights were set as 0.5 and 0.5. The results are presented in Table 3.

From Table 3 we can conclude that the topic based method is more effective than the traditional method for at least the following 2 aspects. First, 50% of the Top 10 institutions resulted by the proposed topic based method are the same as US News Ranking, much higher than the 20% of those by the traditional method. Second, for those toppest institutions, 2 of the Top 3 institutions resulted by the proposed topic based method are also in the Top 3 in the US News Ranking, and the Carnegie Mellon University is also ranked in Top 10. While for traditional method, none of the Top 3 institutions in the US News Ranking appears in the Top 3 list, and only Stanford University appears in the Top 10 list.

---

3. tech.163.com/photoview/6PGI0009/13525.html
Table 3 Comparison of the 3 Rankings

<table>
<thead>
<tr>
<th>RANKING</th>
<th>USNEWS</th>
<th>TOPIC BASED</th>
<th>TRADITIONAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Stanford University</td>
<td>MIT</td>
<td>University of Texas at Austin</td>
</tr>
<tr>
<td>2</td>
<td>Carnegie Mellon University</td>
<td>Stanford University</td>
<td>Nanyang Technological University</td>
</tr>
<tr>
<td>3</td>
<td>MIT</td>
<td>Nanyang Technological University</td>
<td>Sydney University of Science and Technology</td>
</tr>
<tr>
<td>4</td>
<td>UC Berkeley</td>
<td>Microsoft</td>
<td>Washington University</td>
</tr>
<tr>
<td>5</td>
<td>University of Washington</td>
<td>Chinese Academy of Sciences</td>
<td>National Taiwan University of Science and Technology</td>
</tr>
<tr>
<td>6</td>
<td>Georgia Institute of Technology</td>
<td>Texas State University, Austin</td>
<td>Stanford University</td>
</tr>
<tr>
<td>7</td>
<td>University of Illinois at Urbana - Champaign</td>
<td>Carnegie Mellon University</td>
<td>Valencia University of Technology</td>
</tr>
<tr>
<td>8</td>
<td>Texas State University, Austin</td>
<td>University of London</td>
<td>Carlos III University of Madrid</td>
</tr>
<tr>
<td>9</td>
<td>Cornell University</td>
<td>Castilla-La Mancha University</td>
<td>Shanghai Jiaotong University</td>
</tr>
<tr>
<td>10</td>
<td>University of California at Los Angeles</td>
<td>University of Illinois at Urbana - Champaign</td>
<td>Indian Institute of Technology</td>
</tr>
</tbody>
</table>

Finally, by setting $\tau = 0.15$, $p = 10\%$, then MIT, Stanford University, Nanyang Technological University, Microsoft, Chinese Academy of Sciences, Texas State University-Austin, and Carnegie Mellon University are leading the research on AI. While Beijing University of Posts and Telecommunications, Northeastern University, Wuhan University, Graz Technical University, Paderborn University, Malaysia University of Technology and many other institutions still need to improve their research on AI.

Conclusion

This paper presents a method of evaluating the competitiveness of research institutions based on research topic distribution. The method uses the LDA topic model to obtain paper-topic distribution matrix to objectively assign the academic impact of papers (such as times of citation) to research topics. Then the method calculates the competitiveness of each research institution on each research topic with the help of institution-paper matrix. Finally, the competitiveness and the research strength and/or weakness of the institutions are defined and characterized. Case study shows that the method can lead to an objective and effective evaluation of the research competitiveness of given research institutions on given research field.

It is to be noted that the proposed method formally defines the process competitiveness evaluation, i.e., topic analysis, impact allocation and competitiveness measurement. Many adaptations can be easily applied in practice. For example, one can integrate the process of impact allocation with expert preference by multiplying $\Theta^T$ with a weight vector. In the meanwhile, more indicators such as centrality of a paper in the citation network can also be considered for paper impact characterization. Besides, since LDA model can handle topics on multi-granularity, by considering different research bodies (e.g., countries, institutions or researchers), the method can be easily used to evaluate the competitiveness of research bodies and topics on various levels of details. Our future work will consider the mentioned adaptations for a more effective competitiveness evaluation. Other work like mining patterns among strengths and/or weaknesses may also be considered so that the relationships among topics can be further understood.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant (No. 7160325); the Young Talent-Field Frontier Project of Wuhan Documentation and Information Center, Chinese Academy of Sciences.
References
Open access papers: their growth over time and from different countries, and their citations

Grant Lewison
grant.lewison@kcl.ac.uk
King's College London, Cancer Studies, Guy's Hospital, London SE1 9RT (UK)

Abstract
The Web of Science now records the Open Access (OA) status of papers that it covers. Papers with OA = yes have grown rapidly since 2000 and the proportion has reached 18% of the total in biomedical research but only 4% in chemistry. Countries also vary greatly in their preference for publishing in OA journals: Brazil tops the list in four fields but not in physics where the Ukraine and Russia use OA the most among the 30 top publishers in 2015. Contrary to popular belief that OA papers would receive more citations, this appears not to be true, probably because the top-cited journals are not OA, some OA journals may have lower peer-review and editorial standards, and because in some countries (e.g., Brazil, Spain) many OA journals are not in English. However, there is some evidence that OA papers may be attracting more citations from researchers in lower-middle income and low-income countries.

Conference Topics
journals, databases and electronic publications; country-level studies; citation analysis

Keywords
open access; countries; major fields

Introduction
Whereas the traditional publishing model was that it did not cost researchers to publish their papers in peer-reviewed journals, there is now an increasing tendency to ask most of them to pay for this privilege (ones in low-income countries may be excused), but in return to provide free access online to their papers to everyone. This is the "gold" open access (OA) model. There is also the "green" OA model, where authors are permitted to self-archive their papers in an institutional repository while access through the journal is not available. Other models include "delayed" OA, whereby access to the content of the journal is freely available after an embargo period, typically 12 or 24 months, and "hybrid" OA, whereby some articles are freely available immediately because the author(s) have paid a fee: Scientometrics is in this category. Not all OA journals charge authors a fee: some are subsidised by academic institutions, learned societies or government agencies.

The question of whether OA papers receive more citations than non-OA ones is inevitably rather vexed because there are so many factors that can influence citation scores, and all of them need to be taken into account if the influence of just one of them, OA or not OA, is to be determined. This factor may also influence other supposedly independent variables, such as the decision of the authors on which journal to select for their paper based on their opinion on its likelihood of acceptance and perhaps its cost of publication. Thus some studies (Eysenbach, 2006; Gargouri et al, 2010; Sotudeh et al., 2015; Wang et al., 2015) show OA papers having a citation advantage, others (Davis et al, 2008; Davis, 2011; Torres-Salinas et al., 2016) the reverse, and some claiming that there is little difference (Moed, 2007; Björk & Solomon, 2012; Wray, 2016). A contributing factor seems to be whether the OA journals charge the authors: those that do appear to have a citation advantage (Solomon et al., 2013. There is also some dispute over whether OA leads to citations occurring earlier (Moed, 2007; Davis et al., 2008).
We wished to investigate how the practice of OA had grown over the years, to see which scientific fields embraced it most, and which countries had taken to OA with the greatest enthusiasm. We also hypothesised that low-income countries, whose institutions might not have been able to afford to pay for many subscription journals, would preferentially benefit from OA and that therefore their researchers would have a larger share of the citations to OA papers than of those of non-OA papers.

The Web of Science (WoS) now records whether papers are OA or not, but the positive marking refers only to journals that are completely OA, and not to hybrid ones where authors (usually the minority) may choose to pay for their papers to be OA. We used this marking to distinguish between OA and non-OA papers, and to investigate the temporal evolution of OA in five major fields and in different countries. We also looked at the five-year citation scores of different groups of papers, with particular attention to the countries that were citing them. For this purpose they were grouped into four categories by their gross national income per capita: the boundaries were:

- high income / $12,475
- upper middle / $4035
- lower middle / $1025
- low income

according to the World Bank in 2016 (World Bank, 2016).

**Methodology**

Articles were selected from the WoS for the 17 years, 2000 to 2016, in five major fields: biology & environment (BIOL), biomedical research (BMED), chemistry (CHEM), engineering & technology (ENGR) and physics (PHYS). These were chosen to cover the major areas of science. Chemistry and physics are discrete subject areas, but biomedical research was defined by means of address words and contractions indicative of the subject area (Lewison & Paraje, 2004), for example:

\[ AD=(BIOCH*M \text{ OR } CLIN \text{ OR } GENET \text{ OR } HLTH* \text{ OR } HOSP* \text{ OR } MED \text{ OR } NEURO* \text{ OR } PATHO* \text{ OR } P*EDIAT \text{ OR } PHARM* \text{ OR } SURG) \]

The two other fields were composites, formed of several WoS subject areas:

\[ SU=(AGRICULTURE \text{ OR } BIODIVERSITY \text{ OR } CONSERVATION \text{ OR } ENTOMOLOGY \text{ OR } ENVIRONMENTAL SCIENCES \text{ OR } ECOLGY \text{ OR } EVOLUTIONARY BIOLOGY \text{ OR } FISHERIES \text{ OR } FOOD SCIENCE \text{ OR } TECHNOLOGY \text{ OR } FORESTRY \text{ OR } MARINE \text{ OR } FRESHWATER BIOLOGY \text{ OR } MYCOLOGY \text{ OR } PLANT SCIENCES \text{ OR } REPRODUCTIVE BIOLOGY \text{ OR } ZOOLOGY) = BIOL \]

\[ SU=(AUTOMATION \text{ OR } CONTROL SYSTEMS \text{ OR } CONSTRUCTION \text{ OR } BUILDING TECHNOLOGY \text{ OR } ENERGY \text{ OR } FUELS \text{ OR } ENGINEERING \text{ OR } INSTRUMENTATION \text{ OR } MECHANICS \text{ OR } METALLURGY \text{ OR } METALLURGICAL ENGINEERING \text{ OR } NUCLEAR SCIENCE \text{ OR } TECHNOLOGY \text{ OR } ROBOTICS \text{ OR } TELECOMMUNICATIONS) = ENGR \]

After the papers were selected, the Open Access tab allowed the numbers in OA journals to be seen, and the numbers, year by year, in each field to be tabulated. It was also possible to tabulate the numbers from each country (up to about 190 in total) that were either OA or not-OA.

We selected pairs of countries and major fields, based on the above tabulations, for more detailed analysis of the numbers of citations in the five years following publication and the distribution of the citing papers among countries in those five years. The combinations selected were as follows; they were based on large numbers of OA papers in the country concerned:
The citing countries were characterised by their income levels, see above, and the percentages of citations from each group were tabulated. However, because of international co-authorship of many of these citing papers, the sum of the percentages for these four groups was greater than 100%, and so they were proportionately reduced to sum to this total. We found that many of the OA papers from Brazil were in Portuguese (69%) and those from Spain were in Spanish (26%), so it was unlikely that they would be cited as frequently as the non-OA papers (which proved to be the case) but perhaps more often cited by researchers from the 20 Latin American countries, so this was investigated.

**Results**

*Growth of OA in five fields of science*

There was a remarkable difference, particularly in the last seven years, in the numbers of OA papers in biomedical research, which reached over 18% in 2016, compared with the non-biological fields. Figure 1, below, shows the percentages of papers in OA in the five fields, as moving three-year averages so as to smooth out annual fluctuations.

![Figure 1. Growth of Open Access among papers in five fields processed for the Web of Science: BMED = biomedical research; BIOL = biology & environment; CHEM = chemistry; ENGR = engineering & technology; PHYS = physics. Three-year moving averages.](image)

Biology & environmental science initially led biomedical research, but its rate of increase decreased after 2008, although it still outpaced the three physical science fields, particularly chemistry where only 4% of papers were OA in 2016. Tables 2 to 6 show the positions of the 30 countries in each field with the most papers in 2015 relative to the world average in the four years, 2000, 2005, 2010 and 2015, in terms of percentages of OA papers, and the ISO2 codes for the countries are listed in Table 1.
Table 1. Countries with outputs in the top 30 in one or more fields in 2015, with ISO2 codes.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>AR</td>
<td>Finland</td>
<td>FI</td>
<td>Netherlands</td>
<td>NL</td>
<td>South Korea</td>
<td>KR</td>
</tr>
<tr>
<td>Australia</td>
<td>AU</td>
<td>France</td>
<td>FR</td>
<td>New Zealand</td>
<td>NZ</td>
<td>Spain</td>
<td>ES</td>
</tr>
<tr>
<td>Austria</td>
<td>AT</td>
<td>Germany</td>
<td>DE</td>
<td>Norway</td>
<td>NO</td>
<td>Sweden</td>
<td>SE</td>
</tr>
<tr>
<td>Belgium</td>
<td>BE</td>
<td>India</td>
<td>IN</td>
<td>Poland</td>
<td>PL</td>
<td>Switzerland</td>
<td>CH</td>
</tr>
<tr>
<td>Brazil</td>
<td>BR</td>
<td>Iran</td>
<td>IR</td>
<td>Portugal</td>
<td>PT</td>
<td>Taiwan</td>
<td>TW</td>
</tr>
<tr>
<td>Canada</td>
<td>CA</td>
<td>Israel</td>
<td>IL</td>
<td>Romania</td>
<td>RO</td>
<td>Turkey</td>
<td>TR</td>
</tr>
<tr>
<td>China</td>
<td>CN</td>
<td>Italy</td>
<td>IT</td>
<td>Russia</td>
<td>RU</td>
<td>Ukraine</td>
<td>UA</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>CZ</td>
<td>Japan</td>
<td>JP</td>
<td>Saudi Arabia</td>
<td>SA</td>
<td>United Kingdom</td>
<td>UK</td>
</tr>
<tr>
<td>Denmark</td>
<td>DK</td>
<td>Malaysia</td>
<td>MY</td>
<td>Singapore</td>
<td>SG</td>
<td>United States</td>
<td>US</td>
</tr>
<tr>
<td>Egypt</td>
<td>EG</td>
<td>Mexico</td>
<td>MX</td>
<td>South Africa</td>
<td>ZA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Ratios of countries' use of OA for their Web of Science (WoS) papers in biomedical research in four years to the world average. For country codes, see Table 1.

Ratios > 2.0 in 14pt bold; > 1.41 in 12 pt bold; < 0.71 in 12 pt italics; < 0.5 in 10 pt italics.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>17.1</td>
<td>6.61</td>
<td>3.70</td>
<td>1.86</td>
<td>7.32</td>
<td>0.56</td>
<td>0.87</td>
<td>1.03</td>
<td>1.08</td>
</tr>
<tr>
<td>IN</td>
<td>4.80</td>
<td>2.16</td>
<td>1.71</td>
<td>1.11</td>
<td>2.44</td>
<td>0.66</td>
<td>0.93</td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td>PL</td>
<td>5.00</td>
<td>1.65</td>
<td>1.73</td>
<td>1.23</td>
<td>2.40</td>
<td>0.84</td>
<td>0.83</td>
<td>0.90</td>
<td>0.96</td>
</tr>
<tr>
<td>CN</td>
<td>5.60</td>
<td>1.54</td>
<td>1.10</td>
<td>1.18</td>
<td>2.36</td>
<td>0.47</td>
<td>0.81</td>
<td>1.00</td>
<td>1.09</td>
</tr>
<tr>
<td>CZ</td>
<td>2.33</td>
<td>2.53</td>
<td>1.06</td>
<td>1.00</td>
<td>1.73</td>
<td>0.26</td>
<td>0.53</td>
<td>1.25</td>
<td>1.28</td>
</tr>
<tr>
<td>ES</td>
<td>1.63</td>
<td>1.04</td>
<td>1.01</td>
<td>1.07</td>
<td>1.19</td>
<td>0.46</td>
<td>0.59</td>
<td>0.78</td>
<td>1.46</td>
</tr>
<tr>
<td>NO</td>
<td>0.34</td>
<td>1.28</td>
<td>1.34</td>
<td>1.31</td>
<td>1.07</td>
<td>0.53</td>
<td>0.83</td>
<td>0.87</td>
<td>0.97</td>
</tr>
<tr>
<td>SE</td>
<td>0.77</td>
<td>1.07</td>
<td>1.23</td>
<td>1.20</td>
<td>1.07</td>
<td>0.44</td>
<td>0.76</td>
<td>0.90</td>
<td>1.03</td>
</tr>
<tr>
<td>IR</td>
<td>0.58</td>
<td>0.88</td>
<td>1.67</td>
<td>1.02</td>
<td>1.04</td>
<td>0.35</td>
<td>0.75</td>
<td>0.93</td>
<td>1.05</td>
</tr>
<tr>
<td>PT</td>
<td>0.58</td>
<td>0.94</td>
<td>1.47</td>
<td>1.07</td>
<td>1.01</td>
<td>0.39</td>
<td>0.77</td>
<td>0.74</td>
<td>1.01</td>
</tr>
<tr>
<td>IT</td>
<td>1.02</td>
<td>1.07</td>
<td>0.81</td>
<td>0.89</td>
<td>0.95</td>
<td>0.64</td>
<td>0.70</td>
<td>0.58</td>
<td>0.92</td>
</tr>
<tr>
<td>DK</td>
<td>0.59</td>
<td>1.09</td>
<td>1.05</td>
<td>1.05</td>
<td>0.95</td>
<td>0.45</td>
<td>0.77</td>
<td>0.80</td>
<td>0.83</td>
</tr>
<tr>
<td>TR</td>
<td>0.81</td>
<td>1.06</td>
<td>1.11</td>
<td>0.73</td>
<td>0.93</td>
<td>0.28</td>
<td>0.93</td>
<td>0.74</td>
<td>0.88</td>
</tr>
<tr>
<td>CH</td>
<td>0.37</td>
<td>1.07</td>
<td>1.13</td>
<td>1.13</td>
<td>0.93</td>
<td>0.40</td>
<td>0.68</td>
<td>0.47</td>
<td>0.73</td>
</tr>
<tr>
<td>FI</td>
<td>0.62</td>
<td>1.01</td>
<td>0.92</td>
<td>1.05</td>
<td>0.90</td>
<td>0.30</td>
<td>0.48</td>
<td>0.53</td>
<td>0.88</td>
</tr>
</tbody>
</table>
### Table 3. Ratios of countries’ use of OA for their Web of Science (WoS) papers in biology & environmental research in four years to the world average. For country codes, see Table 1.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>14.3</td>
<td>9.19</td>
<td>6.15</td>
<td>3.28</td>
<td>8.23</td>
<td>BE</td>
<td>0.44</td>
<td>0.47</td>
<td>0.84</td>
<td>0.83</td>
<td>0.65</td>
</tr>
<tr>
<td>PL</td>
<td>3.64</td>
<td>2.21</td>
<td>1.06</td>
<td>1.27</td>
<td>2.04</td>
<td>NL</td>
<td>0.43</td>
<td>0.53</td>
<td>0.65</td>
<td>0.97</td>
<td>0.64</td>
</tr>
<tr>
<td>CZ</td>
<td>0.43</td>
<td>3.62</td>
<td>2.19</td>
<td>1.51</td>
<td>1.94</td>
<td>ES</td>
<td>0.48</td>
<td>0.45</td>
<td>0.87</td>
<td>0.77</td>
<td>0.64</td>
</tr>
<tr>
<td>MX</td>
<td>1.90</td>
<td>1.02</td>
<td>1.52</td>
<td>1.59</td>
<td>1.51</td>
<td>UK</td>
<td>0.25</td>
<td>0.36</td>
<td>0.57</td>
<td>0.92</td>
<td>0.52</td>
</tr>
<tr>
<td>IT</td>
<td>0.40</td>
<td>3.10</td>
<td>0.81</td>
<td>1.02</td>
<td>1.33</td>
<td>TW</td>
<td>0.24</td>
<td>0.31</td>
<td>0.40</td>
<td>1.10</td>
<td>0.51</td>
</tr>
<tr>
<td>FR</td>
<td>2.39</td>
<td>0.56</td>
<td>0.82</td>
<td>0.91</td>
<td>1.17</td>
<td>CA</td>
<td>0.18</td>
<td>0.37</td>
<td>0.52</td>
<td>0.85</td>
<td>0.48</td>
</tr>
<tr>
<td>AR</td>
<td>1.12</td>
<td>0.86</td>
<td>1.31</td>
<td>1.20</td>
<td>1.12</td>
<td>ZA</td>
<td>0.19</td>
<td>0.20</td>
<td>0.50</td>
<td>0.98</td>
<td>0.47</td>
</tr>
<tr>
<td>TR</td>
<td>0.71</td>
<td>0.81</td>
<td>1.01</td>
<td>0.96</td>
<td>0.87</td>
<td>RU</td>
<td>0.25</td>
<td>0.35</td>
<td>0.60</td>
<td>0.67</td>
<td>0.47</td>
</tr>
<tr>
<td>DE</td>
<td>0.84</td>
<td>0.56</td>
<td>0.80</td>
<td>0.96</td>
<td>0.79</td>
<td>IR</td>
<td>0.00</td>
<td>0.35</td>
<td>0.73</td>
<td>0.70</td>
<td>0.44</td>
</tr>
<tr>
<td>NO</td>
<td>0.61</td>
<td>0.73</td>
<td>0.72</td>
<td>0.93</td>
<td>0.75</td>
<td>CN</td>
<td>0.24</td>
<td>0.23</td>
<td>0.46</td>
<td>0.77</td>
<td>0.42</td>
</tr>
<tr>
<td>SE</td>
<td>0.46</td>
<td>0.54</td>
<td>0.78</td>
<td>1.10</td>
<td>0.72</td>
<td>KR</td>
<td>0.26</td>
<td>0.12</td>
<td>0.43</td>
<td>0.71</td>
<td>0.38</td>
</tr>
<tr>
<td>DK</td>
<td>0.33</td>
<td>0.84</td>
<td>0.78</td>
<td>0.91</td>
<td>0.71</td>
<td>AU</td>
<td>0.03</td>
<td>0.31</td>
<td>0.37</td>
<td>0.77</td>
<td>0.37</td>
</tr>
<tr>
<td>PT</td>
<td>0.87</td>
<td>0.62</td>
<td>0.65</td>
<td>0.67</td>
<td>0.70</td>
<td>JP</td>
<td>0.09</td>
<td>0.39</td>
<td>0.38</td>
<td>0.56</td>
<td>0.35</td>
</tr>
<tr>
<td>CH</td>
<td>0.40</td>
<td>0.62</td>
<td>0.68</td>
<td>0.97</td>
<td>0.67</td>
<td>NZ</td>
<td>0.04</td>
<td>0.21</td>
<td>0.45</td>
<td>0.61</td>
<td>0.33</td>
</tr>
<tr>
<td>US</td>
<td>0.57</td>
<td>0.56</td>
<td>0.63</td>
<td>0.82</td>
<td>0.65</td>
<td>IN</td>
<td>0.16</td>
<td>0.28</td>
<td>0.32</td>
<td>0.43</td>
<td>0.30</td>
</tr>
</tbody>
</table>

### Table 4. Ratios of countries’ use of OA for their Web of Science (WoS) papers in chemistry research in four years to the world average. For country codes, see Table 1.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>12.1</td>
<td>13.5</td>
<td>7.61</td>
<td>2.51</td>
<td>8.93</td>
<td>IT</td>
<td>0.46</td>
<td>0.45</td>
<td>0.75</td>
<td>1.44</td>
<td>0.78</td>
</tr>
<tr>
<td>SA</td>
<td>3.85</td>
<td>3.62</td>
<td>8.22</td>
<td>1.78</td>
<td>4.37</td>
<td>CA</td>
<td>0.71</td>
<td>0.89</td>
<td>0.54</td>
<td>0.88</td>
<td>0.75</td>
</tr>
<tr>
<td>EG</td>
<td>4.44</td>
<td>3.52</td>
<td>2.68</td>
<td>2.01</td>
<td>3.16</td>
<td>BE</td>
<td>0.84</td>
<td>0.59</td>
<td>0.44</td>
<td>0.80</td>
<td>0.67</td>
</tr>
<tr>
<td>MY</td>
<td>1.31</td>
<td>0.80</td>
<td>6.01</td>
<td>2.40</td>
<td>2.63</td>
<td>KR</td>
<td>0.07</td>
<td>0.24</td>
<td>0.96</td>
<td>1.34</td>
<td>0.65</td>
</tr>
<tr>
<td>RO</td>
<td>4.02</td>
<td>1.94</td>
<td>1.28</td>
<td>1.24</td>
<td>2.12</td>
<td>UK</td>
<td>0.53</td>
<td>0.57</td>
<td>0.47</td>
<td>0.86</td>
<td>0.61</td>
</tr>
<tr>
<td>IR</td>
<td>3.16</td>
<td>2.01</td>
<td>1.31</td>
<td>0.54</td>
<td>1.75</td>
<td>DE</td>
<td>0.28</td>
<td>0.60</td>
<td>0.66</td>
<td>0.85</td>
<td>0.60</td>
</tr>
<tr>
<td>TR</td>
<td>0.87</td>
<td>3.10</td>
<td>1.33</td>
<td>1.00</td>
<td>1.57</td>
<td>CN</td>
<td>0.27</td>
<td>0.50</td>
<td>0.61</td>
<td>0.91</td>
<td>0.57</td>
</tr>
<tr>
<td>MX</td>
<td>0.86</td>
<td>0.96</td>
<td>1.91</td>
<td>2.46</td>
<td>1.55</td>
<td>FR</td>
<td>0.57</td>
<td>0.59</td>
<td>0.50</td>
<td>0.64</td>
<td>0.57</td>
</tr>
<tr>
<td>CZ</td>
<td>2.04</td>
<td>1.06</td>
<td>1.08</td>
<td>1.53</td>
<td>1.42</td>
<td>NL</td>
<td>0.24</td>
<td>0.34</td>
<td>0.69</td>
<td>0.72</td>
<td>0.50</td>
</tr>
<tr>
<td>TW</td>
<td>0.13</td>
<td>0.18</td>
<td>1.84</td>
<td>2.30</td>
<td>1.11</td>
<td>SE</td>
<td>0.19</td>
<td>0.20</td>
<td>0.65</td>
<td>0.89</td>
<td>0.48</td>
</tr>
<tr>
<td>PL</td>
<td>0.17</td>
<td>1.17</td>
<td>0.73</td>
<td>2.32</td>
<td>1.10</td>
<td>CH</td>
<td>0.16</td>
<td>0.43</td>
<td>0.49</td>
<td>0.57</td>
<td>0.41</td>
</tr>
<tr>
<td>ES</td>
<td>0.98</td>
<td>0.74</td>
<td>1.08</td>
<td>1.49</td>
<td>1.08</td>
<td>US</td>
<td>0.34</td>
<td>0.25</td>
<td>0.33</td>
<td>0.65</td>
<td>0.39</td>
</tr>
<tr>
<td>PT</td>
<td>0.48</td>
<td>1.17</td>
<td>0.86</td>
<td>0.96</td>
<td>0.87</td>
<td>RU</td>
<td>0.50</td>
<td>0.46</td>
<td>0.13</td>
<td>0.34</td>
<td>0.36</td>
</tr>
<tr>
<td>AU</td>
<td>1.16</td>
<td>0.91</td>
<td>0.57</td>
<td>0.86</td>
<td>0.82</td>
<td>JP</td>
<td>0.10</td>
<td>0.23</td>
<td>0.41</td>
<td>0.65</td>
<td>0.35</td>
</tr>
<tr>
<td>IN</td>
<td>0.30</td>
<td>0.66</td>
<td>1.88</td>
<td>0.44</td>
<td>0.82</td>
<td>SG</td>
<td>0.00</td>
<td>0.23</td>
<td>0.23</td>
<td>0.53</td>
<td>0.25</td>
</tr>
</tbody>
</table>
### Table 5. Ratios of countries' use of OA for their Web of Science (WoS) papers in engineering & technology research in four years to the world average. For country codes, see Table 1.

<table>
<thead>
<tr>
<th>ISO</th>
<th>2000</th>
<th>2005</th>
<th>2010</th>
<th>2015</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>16.4</td>
<td>4.13</td>
<td>2.24</td>
<td>1.18</td>
<td>5.98</td>
</tr>
<tr>
<td>ES</td>
<td>3.78</td>
<td>4.86</td>
<td>2.33</td>
<td>1.35</td>
<td>3.08</td>
</tr>
<tr>
<td>MX</td>
<td>2.48</td>
<td>1.96</td>
<td>3.04</td>
<td>1.44</td>
<td>2.23</td>
</tr>
<tr>
<td>KR</td>
<td>2.59</td>
<td>2.30</td>
<td>1.27</td>
<td>1.58</td>
<td>1.94</td>
</tr>
<tr>
<td>PL</td>
<td>0.61</td>
<td>2.17</td>
<td>0.65</td>
<td>3.33</td>
<td>1.69</td>
</tr>
<tr>
<td>JP</td>
<td>2.44</td>
<td>2.43</td>
<td>0.52</td>
<td>0.58</td>
<td>1.49</td>
</tr>
<tr>
<td>IN</td>
<td>2.61</td>
<td>1.14</td>
<td>0.77</td>
<td>0.56</td>
<td>1.27</td>
</tr>
<tr>
<td>SA</td>
<td>1.38</td>
<td>0.86</td>
<td>1.40</td>
<td>1.22</td>
<td>1.22</td>
</tr>
<tr>
<td>FR</td>
<td>1.88</td>
<td>1.15</td>
<td>1.04</td>
<td>0.50</td>
<td>1.14</td>
</tr>
<tr>
<td>EG</td>
<td>0.91</td>
<td>1.11</td>
<td>1.14</td>
<td>0.87</td>
<td>1.01</td>
</tr>
<tr>
<td>TR</td>
<td>0.69</td>
<td>1.43</td>
<td>1.04</td>
<td>0.71</td>
<td>0.97</td>
</tr>
<tr>
<td>TW</td>
<td>0.29</td>
<td>0.41</td>
<td>1.12</td>
<td>1.41</td>
<td>0.81</td>
</tr>
<tr>
<td>MY</td>
<td>0.00</td>
<td>0.36</td>
<td>1.54</td>
<td>1.33</td>
<td>0.81</td>
</tr>
<tr>
<td>FI</td>
<td>0.60</td>
<td>0.51</td>
<td>1.30</td>
<td>0.69</td>
<td>0.78</td>
</tr>
<tr>
<td>BE</td>
<td>0.94</td>
<td>0.43</td>
<td>1.07</td>
<td>0.63</td>
<td>0.77</td>
</tr>
</tbody>
</table>

### Table 6. Ratios of countries' use of OA for their Web of Science (WoS) papers in physics research in four years to the world average. For country codes, see Table 1.

<table>
<thead>
<tr>
<th>ISO</th>
<th>2000</th>
<th>2005</th>
<th>2010</th>
<th>2015</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>RU</td>
<td>4.18</td>
<td>4.22</td>
<td>2.74</td>
<td>1.43</td>
<td>3.14</td>
</tr>
<tr>
<td>IN</td>
<td>5.55</td>
<td>3.16</td>
<td>1.62</td>
<td>1.04</td>
<td>2.84</td>
</tr>
<tr>
<td>AT</td>
<td>0.00</td>
<td>1.29</td>
<td>1.59</td>
<td>1.84</td>
<td>1.18</td>
</tr>
<tr>
<td>CH</td>
<td>0.31</td>
<td>0.64</td>
<td>1.13</td>
<td>2.33</td>
<td>1.10</td>
</tr>
<tr>
<td>JP</td>
<td>1.19</td>
<td>1.04</td>
<td>0.86</td>
<td>1.15</td>
<td>1.06</td>
</tr>
<tr>
<td>PL</td>
<td>0.58</td>
<td>0.75</td>
<td>0.64</td>
<td>1.95</td>
<td>0.98</td>
</tr>
<tr>
<td>SE</td>
<td>0.46</td>
<td>0.93</td>
<td>1.04</td>
<td>1.42</td>
<td>0.96</td>
</tr>
<tr>
<td>DE</td>
<td>0.28</td>
<td>0.85</td>
<td>1.32</td>
<td>1.39</td>
<td>0.96</td>
</tr>
<tr>
<td>IL</td>
<td>0.40</td>
<td>0.85</td>
<td>0.82</td>
<td>1.67</td>
<td>0.94</td>
</tr>
<tr>
<td>UK</td>
<td>0.21</td>
<td>0.81</td>
<td>0.97</td>
<td>1.57</td>
<td>0.89</td>
</tr>
<tr>
<td>IT</td>
<td>0.17</td>
<td>0.64</td>
<td>0.99</td>
<td>1.69</td>
<td>0.87</td>
</tr>
<tr>
<td>ES</td>
<td>0.14</td>
<td>0.54</td>
<td>1.08</td>
<td>1.72</td>
<td>0.87</td>
</tr>
<tr>
<td>MX</td>
<td>0.22</td>
<td>0.74</td>
<td>0.64</td>
<td>1.68</td>
<td>0.82</td>
</tr>
<tr>
<td>IR</td>
<td>1.28</td>
<td>0.48</td>
<td>0.74</td>
<td>0.75</td>
<td>0.81</td>
</tr>
</tbody>
</table>

The tables show that for most countries there has been a regression to the mean, so that countries that either published many more (or many fewer) OA papers than the world mean...
in 2000 had become closer to the world average in 2015. One might have expected that, since it often costs researchers to publish in OA journals, that richer countries would have chosen OA journals more than poor countries, but this is not the case. The USA has been below the world average in all five major fields, and Switzerland (CH) in all except for physics (where the presence of CERN may have made a difference); on the other hand Brazil has been well above average in all fields except for physics, and India has been above average in physics, biomedical research and engineering, and only a bit below average in chemistry.

Citations to OA and non-OA papers
We investigated the citations to papers from five combinations of countries and major fields. The first of these was Brazilian papers in biology & environmental science. The mean five-year citation scores for OA and non-OA papers from 2006-11 were 2.2 and 6.1, showing that the non-OA papers were cited much more than OA ones. This was not surprising because so many of the papers were in Portuguese. A consequence was that 71% of the citations to the OA papers were from Latin American countries, including Brazil, but only 35% of the citations to the non-OA papers.

The second set of papers was biomedical papers from Sweden. These papers were all in English, and the five-year citation scores were very similar, 15.2 for the OA papers and 15.3 for the non-OA ones. An analysis of the citing countries in the five years broken down by income groups showed, as expected, that over 85% of all citations came from high-income countries. However, the lower-middle income countries accounted for 2.8% of the citations to the OA papers but only 1.8% of the citations to non-OA papers. There was an even bigger ratio for the low income countries: they provided 1.1% of the cites to the OA papers but only 0.3% of the cites to non-OA papers. So the OA papers were attracting more interest (cites) from the poorer countries than the non-OA ones, although these are only a tiny proportion of the total numbers of cites.

The third combination was chemistry papers from Germany. Here the non-OA papers again had an advantage over the OA ones in mean five-year citation scores, 12.6 cites to 8.7. And, as with the Swedish biomedical papers, the OA papers received a higher percentage of their cites from lower-medium income countries (7.0%) than did the non-OA ones (4.3%), and there was a bigger ratio for the low-income countries (0.2% compared with 0.034%). Once again, these represent a small minority of the cites, but OA is benefiting these poorer countries.

The fourth combination, Spanish papers in engineering & technology, showed a bigger citation advantage for the non-OA papers as 26% of the OA papers were in Spanish. The actual mean cite counts were 8.6 and 3.5 citations in five years. The lower-medium- and low-income countries did not benefit from OA, but the Latin American countries did, with 6.7% of citations compared with 4.9% for the non-OA papers.

Finally the fifth combination, of Russian physics papers was analysed. The citation scores were quite low, even though all the papers were in English: 5.4 for non-OA papers and 3.3 for the OA ones. But there were much more dramatic differences in the citing country percentages: the upper-middle countries (which include Russia) accounted for 45.6% of the OA citations compared with 29.1% for the non-OA papers' citations, and there was a small benefit to the lower-middle income countries as well (5.4% of cites compared with 5.0% for the non-OA papers).
Discussion
These are only samples, but they appear fairly consistent in showing that the non-OA papers enjoy a citation advantage over the OA ones. Because there are many other factors that can influence citation scores, this cannot be regarded as a definitive result, but it seems unlikely that OA papers enjoy a citation advantage by virtue simply of being OA. However, there is a very small but fairly consistent advantage for the lower-middle and low income countries in the percentage of citations that they provide.

A more detailed study would examine the relative citation advantages enjoyed by OA papers in hybrid journals. However, since OA publication is increasingly being required by a number of major funding bodies, such as the US National Institutes of Health, and the Wellcome Trust in the UK, and also by many universities, the division between OA and non-OA papers may depend on whether the papers have been externally funded after a rigorous peer-review. This will inevitably mean that funded papers will be more likely to be OA if they are in a hybrid journal, and we know that such papers tend to be of higher quality because of the acceptance of the research that they describe by a prestigious funder. This means that they will be more highly cited than papers without explicit contestable funding, whether or not they are OA.

References


Pisoschi AM & Pisoschi CG. (2016). Is open access the solution to increase the impact of scientific journals? Scientometrics, 109, 1075-1095.  


The Next Generation (plus One): An Analysis of Doctoral Students’ Academic Fertility Using a Novel Approach for Identifying Advisors

April 30, 2017

Abstract
Scientific communities reproduce themselves by allowing senior researchers to train young scholars, in particular through the training of doctoral students. The primary purpose of this paper is to contribute to a better understanding of the self-reproduction of science by analyzing the probability of doctoral students to become advisors themselves. We present a novel approach employing machine learning techniques to identify advisors among (frequent) co-authors in doctoral students’ publications. This approach enabled us to construct a large-scale dataset of German doctoral graduates in applied physics and electrical engineering from 1975 to 2005. Keywords: self-reproduction of science, advisor identification, PhD training, advisor affects, academic careers, peer effects

Acknowledgement: This work was funded by the German Federal Ministry of Education and Research (BMBF) in its program „Forschung zu den Karrierebedingungen und Karriereentwicklungen des Wissenschaftlichen Nachwuchses (FoWiN)” under Grant 16FWN001.

1 Introduction: The production of scientists by means of scientists

Publications are the main output that scientists produce. In publications new results are described and disseminated to other researchers. Publications thus enable the cumulative nature of science, where each generation of scientists stands, as Isaac Newton famously phrased it, “on the shoulder of giants”. However, it is implicit in Newton’s metaphor that publications alone are not sufficient to keep science functioning. A second output is required to do so: a new generations of scientists willing and able to climb atop their predecessors’ shoulders.

Science is a self-reproducing system, and the training of doctoral students is a key step in the process of re-production. New members of scientific communities are spawned within these communities. Specifically, they are trained by existing
members of the community. This is most clearly observed in doctoral training: only established researchers - normally those who hold professorships - have the privilege to graduate doctoral students, i.e., admit new members to the community.

The relevance of producing doctoral students is recognized in the conventions by which reputation is allocated among researchers (e.g., in the notion that person X is “advisor Y’s student”, and also in disciplinary genealogies such as the Mathematics Genealogy Project). It is also a relevant parameter in many performance-based management systems set up by universities and national governments. However, only a fraction of all doctoral students will subsequently become advisors and produce doctoral students themselves, and it is not obvious, for example, whether successful self-reproduction is primarily due to having many offspring or more dependent on offspring quality. To learn more about the self-reproduction of science we therefore need to better understand the determinants of producing “fertile” students, i.e., those of today’s doctoral students that tomorrow will train the next generation of students, and also how these relate to student, advisor and organization characteristics.

The primary purpose of this paper is to contribute to a better understanding of the self-reproduction of science by analyzing the probability of doctoral students to become advisors themselves. To do so, we constructed an original dataset containing more than 20,000 matched student-advisor pairs for German applied physics in the time period from 1975 to 2005. Advisor information is not part of the standard bibliometric data for doctoral dissertations, nor is it otherwise collected and published in systematic form. Prior work therefore has often relied on ad hoc approaches to identify advisor-student pairs. As a second contribution of the present paper, we provide a detailed account of the more systematic new approach we developed. It employs a novel matching algorithm based on publications co-authored by doctoral students and their advisors.

The remainder of this paper is structured as follows. Section 2 develops the theoretical background of the subsequent analysis. It also formulates testable hypotheses. In Section 3 we outline our approach to identify doctoral advisors based on dissertation and co-publication data. Section 4 presents the results of analyzing the dataset developed on the basis of this approach. Section 5 concludes.

2 Theory and hypotheses

2.1 Socialization effects of young researchers

Doctoral training prepares junior researchers for an academic career by enabling them to independently advance the frontier of knowledge. Young researchers need to acquire a broad range of knowledge and skills to perform successfully in academia. These range from mastering the intricate details of their field to in-

\footnote{See: https://genealogy.math.ndsu.nodak.edu/, or for a comparable project in Germany: https://www.bibsonomy.org/persons.}
sights into research strategies and routines of professional etiquette. The various types of knowledge and skills can be learned in a variety of ways. The institutional setup known as Open Science (Dasgupta et al., 1994) provides researchers with strong incentives to codify and freely communicate knowledge via publications. Significant parts of the required knowledge can therefore be accessed by reviewing the relevant literature. However, there is a broad consensus that not all relevant knowledge can be acquired from the literature. For skills such as the handling of laboratory equipment learning from one’s own experience is crucial, often based on a process of trial and error. In addition, besides access to codified knowledge and own experience, direct face-to-face interaction with others is crucial for becoming a successful researcher (e.g. Collins, 1974; Stephan, 2012). Face-to-face interaction allows doctoral students to access non-codified or tacit knowledge that cannot be found in the published scientific literature.

It also enables vicarious learning from observing the behavior of others as well as the environmental reaction to this behavior (Bandura, 1986).

Professors have a strong position in the German university system, and traditionally the training of individual doctoral students has been in the hands of a single doctoral advisor (the Doktorvater). This suggests an important role of the advisor in shaping the knowledge, skills, and attitudes of doctoral students (Buenstorf and Geissler, 2014). In light of these considerations, it seems plausible that doctoral students benefit by learning from their advisor, and that more can be learned from better-performing advisors. In turn, the acquired knowledge enhances a student’s chances to remain in academic, i.e., to become an advisor himself or herself. It is well-established, however, that only a fraction of all graduated doctoral students pursue academic careers. This hold for various time periods and countries (Stephan, 2012; Waldinger, 2016) and also for leading departments (Conley and Önder, 2014).

Advisors are not only important sources of knowledge for junior researchers, but also important gatekeepers of the scientific community. The status that a given advisor enjoys in the community is an important precondition of their students’ ability to signal their quality. Being advised by a highly reputed researcher and/or co-authoring with them thus acts to certify student quality. Students of more highly reputed advisors may also be able to benefit from an extended Matthew effect (Merton, 1968) resulting in their work getting more attention. Advisors perform even more direct gate-keeping functions in referring students within the community (Baruffaldi et al., 2016) and by helping them secure academic jobs (a precondition for them becoming advisors). How well a given advisor can perform these functions is only indirectly dependent on their own performance. More directly relevant is their reputation in the community (which would be expected to be associated with prior performance), but even more their status and influence in the respective community.

We expect both advisor status and advisor performance to exert beneficial influences on their students’ probability to become advisors themselves, and will try to disentangle them in the empirical analysis. Specifically we predict the following effects on students’ academic fertility:
**Hypothesis 1** Doctoral students whose advisor is better positioned in the network of the respective scientific community have a higher probability to remain active in science and to become advisors themselves.

**Hypothesis 2** Doctoral students whose advisor has superior research output have a higher probability to remain active in science and to become advisors themselves.

### 2.2 Life cycle effects

Their advisors’ age and career stage at the time they worked with them may be another relevant factor in doctoral students’ subsequent career outcomes. It is not obvious, however, how the probability that students remain active in research changes over the career trajectory of their advisors.

On the one hand, we would generally expect that more senior advisors are better able to teach their students. We would also expect more senior advisors to be more established in their community. They should therefore be better positioned to leverage their own contacts to the benefit of their students, for example by arranging for talks and visits to research groups possessing complementary knowledge. In addition, doctoral advisors may switch positions during their career. Such moves would generally be expected from less to more highly reputed universities, suggesting that students advised at later career stages may enjoy a better working environment.

On the other hand, life-cycle models of individual research output (Levin and Stephan, 1991) argue that as researchers get closer to their retirement age, they have weaker incentives to engage in research activities. Instead, they may be more strongly inclined to engage in consulting work and/or decrease their own working hours. In addition, while being more experienced and possibly having a broader overview of their field, more senior researchers may lose touch with the most recent developments in terms of theories and empirical methods. Their research foci may increasingly diverge from those that are currently „hot“, which makes it harder for their students to find attention for their own work. As scientific knowledge tends to decay relatively rapidly, more senior advisors may thus be less well-positioned to teach their students. Furthermore, it is plausible that students of more senior researchers face stronger competition for the advisor’s scarce time and attention. More senior advisors will often have a larger burden of administrative and managerial duties in the research group and the department. With increasing group size, they also have to divide their attention between a larger number of students. It moreover seems plausible that with an increasing „stock“ of prior doctoral students, advisors’ interest in securing the „survival“ of their line of research, which requires students who become advisors of the new generation of students, is increasingly muted.

Based on the prior evidence of mostly adverse effects of seniority on research performance, we expect the second set of arguments (suggesting a decreasing likelihood of student fertility with increasing advisor seniority) to be more relevant than the first one (which suggest an increasing likelihood). We also expect
that the role of advisors’ accumulated stock of doctoral students can be separated from that of advisor career age. These considerations inform a second set of hypotheses:

**Hypothesis 3** *Doctoral students whose advisor has a larger number of prior doctoral students have a reduced probability to remain active in science and to become advisors themselves.*

**Hypothesis 4** *Doctoral students whose advisor has reached a higher career age have a reduced probability to remain active in science and to become advisors themselves.*

### 2.3 Peer effects and competition among doctoral students

Doctoral advisors are certainly not the only relevant source of knowledge and skills that doctoral students draw upon. Relevant face-to-face interaction primarily takes place in the setting of the research group or laboratory, with fellow students and co-workers (Krabel, 2012; Tartari et al., 2014; Hottenrott and Lawson, 2017) being important partners from whom doctoral students can learn. In the German context co-workers in the research group have mostly been selected and/or trained by the student’s own advisor. Accordingly, peer group effects exerted by fellow students and co-workers will generally not be independent from advisor effects. It nonetheless seems plausible that students benefit from being part of a strong cohort of researchers contemporaneously working the same research group or laboratory. Other group members are sources of knowledge and social capital, and they may also provide role models affecting the career choices that a student makes. This suggests a positive peer group effect on individual students’ performance and thus their ability and willingness to embark on an academic careers.

**Hypothesis 5.a** *Doctoral students have a higher probability to remain active in science and to become advisors themselves if they are trained together with larger numbers of contemporary peers having the same advisor who remain in academia and subsequently become doctoral advisors.*

The above argument for positive peer effects does not take into account, however, that doctoral students in the German system need to go through additional career stages before assuming their first faculty positions. To a considerable extent these post-doctoral career stages are taken in the same research group or laboratory in which the individual received her doctoral training, which seems to be associated with superior career outcomes (Bäker, 2015). This implies a competitive relationship among the student and her contemporary peers or ”siblings”, since often only a fraction of the graduated doctoral students can remain in the lab. Students may thus suffer from being trained together with exceptionally well trained ”siblings”, which informs a competing hypothesis on peer impact:
Hypothesis 5.b  Doctoral students have a reduced probability to remain active in science and to become advisors themselves if they are trained together with larger numbers of contemporary peers having the same advisor who remain in academia and subsequently become doctoral advisors.

3 Identifying doctoral advisors: a new approach based on co-authored publications

In this section we outline a new large-scale approach to identify the advisors of doctoral dissertations. This approach exploits information about co-authored journal publications and underlies the empirical analysis presented below. We first present the general approach before discussing its implementation for the present paper. To keep the discussion tractable, only the most salient features of the approach are presented in the main body of the article, whereas further technical details are provided in footnotes as well as in Appendix A.

Data Selection
- In our context we select all natural sciences in the dissertation dataset and applied physics publications

Preprocessing
- Standardizing Names and Affiliations
- WoS name disambiguation in two steps: first step based on affiliation, second step combines same affiliation groups

Record Linkage
- Based on first affiliation of disambiguated researchers

Feature Engineering
- All (German) co-authors are considered as potential advisors
- Features depending on individual PhD-coauthor pairs
- Features standardized on PhD-level

Model Selection
- Models trained on a dataset of laser-scientists (3-fold cross validation)
- Performance is evaluated on separated data

Figure 1: Overview of the data processing and machine learning procedure.

Our approach for identifying doctoral advisors (including the pre-processing stage) is summarized in Figure 1. It is grounded in the assumption that, if doctoral students publish scientific articles out of their thesis they are likely to do this together with their advisor. By tracking co-authors of papers related to a certain doctoral thesis we should be able to pick out the advisor with some certainty. In order to do so we need to identify doctoral students’ papers as a subset of all scientific publications. We are able to solve this task by using additional information on doctoral students that, as best we know, has not been used before (in large-scale analyses). Based on knowing who is a doctoral
student we can identify their publications. Tracking the publication output of doctoral students we can retrieve a full list of all co-authors, in which we expect the advisor to be included. For this group of co-authors we can then calculate several characteristics that we expect to reveal the advisor. By using a training dataset we are able to train a model that predicts the advisor with an evaluated accuracy.

3.1 Dissertation data

Our point of departure is information about the universe of doctoral dissertations completed in Germany. Scholarly interest in dissertation data has been limited, especially when used as individual data and linked to other bibliographic data. However, there is a lot of to learn from these data to understand the production of scientific knowledge (Morichika and Shibayama, 2016). We take our information on doctoral dissertations from the online catalog of the German National Library (Deutsche Nationalbibliothek or DNB for short). Since 1969 German universities have been legally required to deposit copies of all completed doctoral dissertations at the DNB. Universities, in turn, enforce mandatory deposit of doctoral dissertations vis-à-vis their doctoral students. To ensure compliance, the doctoral degree is granted only after the respective student has deposited the required number of copies. Due to this principle of mandatory deposit the DNB holdings of doctoral dissertations, and consequently the information in its catalog, are virtually complete. Identifying doctoral dissertations in the catalog is unambiguous as they are indicated within the catalog’s subsection of works from higher education institutions (Hochschulschriften), along with (non-standardized) information about the issuing university and the year of submission. After eliminating double entries (e.g., due to multiple versions - mostly print and online versions - or editions), a total of 894,086 unique authors of doctoral theses was identified for the time period from 1970 to 2010.

3.2 Publication data

Identifying doctoral advisors on the basis of co-authorships first requires us to identify publications authored by the doctoral students in our sample. We use the Web of Science (WoS) database to obtain these publications. Identifying doctoral students’ publications in the WoS requires solving the namesake problem in the first place. The namesake problem (also known as ambiguity or homonymous problem) arises out of individual scholars who share names and subsequently publish using the same author name (cf. e.g. Smallheiser and Torvik, 2009, for a detailed discussion of the namesake problem). Since WoS does not provide a unique identifier for each individual researcher, author identities have to be established by cleaning and standardizing the WoS data. Solving this problem is nontrivial for several reasons\footnote{This task is best be done by using reliable automated procedures for two reasons: a) number of publications is not feasible for manually processing and b) manual processed data seems to provide unreliable results, especially in the case of common names. Even if authors} and becomes especially challeng-
ing when very common names are considered (c.f. i.e. Smalheiser and Torvik, 2009). Our disambiguation procedure is a somewhat heuristic approach, and assessing its quality is complicated because we try to track publications records of doctoral graduates who frequently leave science after just publishing a few papers. CVs including publications records are frequently not available for these individuals, and using the available ones would likely be biased. Accordingly, we do not have a sample to systematically evaluate the disambiguation. We are nonetheless confident about its quality, which is corroborated by comparing distributions of papers attributed to highly frequent and very rare names (see Appendix A.1).

We decided to first identify individual researchers’ publication œuvres and then link them to the doctoral students3. The applied procedure involves two steps (as comparable to, e.g., Wang et al., 2012; Strotmann et al., 2009; Ferreira et al., 2012): First, we search for clusters of papers sharing author names (surname and first name initial) and affiliation. While the retrieved name-affiliation groups might be sufficient to track doctoral student publications, researchers generally tend to change affiliations during their careers. Therefore we apply a second disambiguation step. By disambiguating name-affiliation groups’ sharing author name (but not affiliation), we aim to identify all publications of individual researchers covering their whole career. This procedure has the advantage that (in the first step) same name-affiliations combinations are more likely to belong to an individual scientist, and, (in the second step) once name-affiliation groups are sorted into distinct individuals more information is available when comparing homonyms over different affiliations (i.e. self-citations are easier to identify if more papers are considered).

Before we divide authors by names and affiliation the WoS requires substantial cleaning and standardization. We first process author names (surname and first initial5) by correcting German ß and ë as well as removing punctuation and whitespaces (the latter is important for German data; cf. Schoen et al., 2014), and assign affiliations to single authors wherever possible5. We identify all German universities by comparing affiliations with a list containing decomposed variations of university names. Within all publications corresponding to one name and affiliation search for distinct clusters which represent one individual scientist active at this affiliation. Within groups, the cluster algorithm assigns authorship to one distinct researcher if at least one of the following crite-

---

3An alternative approach would be the use of an assignment procedure as recommended method Ferreira et al. (2012) and Reijnhoudt et al. (2014), where information on individual scientists is collected at first and subsequently publications are assigned to the known individuals. Our decision against the use of assignment procedures is based on the need to identify all publications attributed to the co-authors as well.

4Only in the latest years full author names are available in the Web of Science publication database.

5Author affiliations can be attributed to a single author in certain cases, i.e. for corresponding authors (for a detailed discussion see Reijnhoudt et al. (2014)). Similar to Reijn-houdt et al. (2014) we considered all affiliations listed in the paper as potential true affiliations of the authors whenever a clear assignment was is possible.
ria is fulfilled: same department is mentioned\(^6\), at least one common co-author (identified by name), authors share second name initial, self-citations between the considered articles can be found, authors use same e-mail address\(^7\), and identical (author) keywords are used. In contrast, author references are identified as different individuals if there is a gap of 10 or more years between articles without any intermittent publications. Different full first names and different second name initials divide publications into separate scientists as well.

Before applying the second step of the disambiguation procedure we exclude all subsets. Subsets emerge when a researcher is using several affiliations. Then it is not clear which affiliation is the main affiliation. We choose to decide for the affiliation most papers are assigned to, all others are deleted from the dataset. The homonym problem to solve now gets at a sample of authors assigned to different affiliations. We still observe some overlaps of one to several papers on different disambiguated name-affiliations groups, in the case of several affiliations mentioned on the paper. This might be the case when researchers change affiliations, but still collaborate with previous colleagues. In the case of overlapping publications two name-affiliation groups are attributed to one single researcher. Non-overlapping distinct name-affiliations are then identified as the same (or different) individuals based on the following criteria: shared co-authors (based on disambiguated name-affiliation groups), self-citations between papers of one group to the other, and shared author keywords. Again, authors are identified as different individuals whenever they have different first names, second initials or publication lags of 10 or more years. As mentioned above the quality of the applied disambiguation procedure is evaluated in Appendix A.1.

3.3 Matching dissertations with WoS authors

To obtain publications related to doctoral dissertations the pre-processed DNB and WoS records have to be linked to each other. The primary name matching procedure performs a simple name comparison of the authors’ surnames and first name initials, which however leads to a high number of false positive matches (as described by D’Angelo et al. (2011)). Therefore we additionally require the graduating university to be equal to the earliest researchers’ affiliation in the WoS data. To further reduce false positive matched pairs we only considered pairs to be true matches if they (a) either refer to a unique match (a doctoral graduate with a unique name in the whole DNB sample matches to only one of the identified scientists in WoS); or (b) some similarity between paper titles and thesis title could be found. Similarity between the titles is computed by using a longest common substring algorithm.\(^8\) We decide to only include those matched

---

\(^6\)The department is tracked from the remaining part of authors affiliations when country and zip codes, cities and universities where extracted. Frequently occurring terms as „dept.“ were eliminated as well. Because of the high variation in different spellings remaining department information was compared by computing a Jaccard-similarity.

\(^7\)E-mail addresses are assigned to individual authors on the basis of highest Jaccard-similarities between all authors and e-mail addresses.

\(^8\)Before we compare titles they are standardized by removing (German and English) stopwords. Removing punctuations and numbers as well as whitespaces. In case several authors
pairs with at least one word (with more than five letters) in common, i.e. „laser“.
As a consequence of these requirements, our approach is rather conservative, and
some author-thesis pairs may not be retrieved because they choose very different
titles. Matched pairs with lags between publications exceeding seven years are
excluded from the sample. Matching quality is evaluated by comparing author
first names whenever available. (Results are reported in Appendix A.2.)

3.4 Co-authors of doctoral papers and feature space construction

We expect doctoral students to publish results of their theses in co-authorship
with their advisor. Subsequently we should be able to identify the advisor among
the student’s co-authors. By using the matched papers of doctoral graduates we
can easily obtain a full list of co-authors. To identify the advisor among these
co-authors we calculate several individual characteristics for each of the co-authors.
These characteristics should describe the unique position and importance among
all co-authors for the respective student. Co-authors who started publishing first
after the first publication of a specific doctoral graduate are excluded from the
dataset. Specifically, we use the following characteristics to identify advisors
within the group of co-authors:

- Times published together (number of papers co-authored together):
  We assume that doctoral students likely publish larger shares of their
  thesis-related papers in co-authorship with their advisors as opposed to
  other individuals.

- Publication lag (time lag between first paper of focal student and first
  paper of co-author):
  We assume advisors be active in scientific publishing a certain time period
  before the PhD graduates. Since we do not have strong priors about the
  lag relation, we include the lag variable in a quadratic function.

- Betweenness centrality within the co-author network:
  We assume advisors to be in a very central position, bridging between all
  of the Ph.Ds. co-authors.

- Burt constraint, based on citations within the co-author network:
  We assume that the person who receives most attention among the co-
  authors (receiving most of the citations) can be interpreted as the person
  with a high impact on the group’s work.

- Number of publications and number of citations:
  We assume co-authors with higher reputation / seniority to be more likely
  the advisor. Because the relationship might be non-linear we include these
  measures in a quadratic function.

match same thesis and vice versa ones with highest score are taken, scores are set in relation
to length of dissertation title.
• Author name position on paper:
  We assume that group leaders are more likely to be named last on the paper. Being named last / or first might capture some information about the role of the specific author.

• Number of dissertations a co-author is mentioned on:
  We assume advisors contributing in various ways to many dissertations. Finding a co-author frequently appearing on other students’ papers might indicate who is the advisor of the considered thesis.

• Degree in co-authors network:
  The network degree in the co-author network provides information on who is connected to how many of the co-authors. Advisors might be on average connected to more co-authors because of their central role in the research group.

• Citations received by PhD papers (number of times previous work of the co-authors is cited by students’ papers):
  We assume that doctoral students tend to follow the research agenda of their advisors, which is reflected by citations received by co-author.

• Same university:
  Co-author is affiliated to the degree-granting university at the same time as the student.

• Co-author is PhD:
  Co-author is identified as holding a doctoral degree. Recently graduated students may not be indicated as holding a doctoral degree in our dataset, while already acting as a post-doc who might have a significant influence on students’ work.

We further include several thesis-related characteristics:

• Number of papers

• Number of co-authors

• Year of submission

Before applying machine learning algorithms, variables have to be standardized because we are interested in identifying one specific individual in the group of co-authors. Some groups have generally high values for some of the attributes (e.g., large numbers of publications) but not for others (e.g., citation rates). We use z-score standardization centered to the mean of all co-authors of the focal doctoral student to take this into account.
3.5 Machine learning algorithms

The final step of our identification approach for doctoral advisors is to use the characteristics of co-authors outlined above to find the doctoral advisor among them. To this end we separate our data in two parts. One is used for training algorithms and identifying best parameter settings. The other is used to evaluate the resulting models on an independent dataset. We train four different algorithms that are standard in the data mining literature: the regularized logistic regression model, support vector machine (SVM), random forests, and ada Boost. All algorithms are available as R packages; for a ed discussion see Bishop (2013).

The regularized logistic regression model performs logistic regression including additional penalty terms to the optimized error function to avoid overfitting by reducing the model’s complexity. However, the penalty parameter needs to be specified. We also need to define a threshold when the estimated probability is high enough to be classified in the true class. Support vector machines have been used i.e. for disambiguation tasks by Wang et al. (2012). SVM uses a hyperplane to separate data by maximizing the distance to the vectors (data points) that are closest to the hyperplane. Those vectors are described as support vectors. To make the observed data linearly separable they are transferred into higher dimensional space. This is done by using a kernel, in our case a radial kernel. To avoid overfitting false classifications can be allowed, which however are penalized. Two parameters can be specified: a penalizing parameter, and parameter gamma specifying the radial basis function kernel. Random forest builds on decision trees. Multiple decision trees are constructed by choosing a random set of features. Decision trees simply split the dataset to achieve best separation of two classes. Combining several splits leads to decisions regions. After training a specific number of trees, all trees are used together to give a majority vote on the class of an object. As parameters the number of randomly drawn features as well as the number of trees need to be specified. The ada Boost algorithm is a boosting technique specified for binary response that has also been used by Han et al. (2004) for author name disambiguation. We use it as a fourth method even though it is also based on decision trees. Different from random Forests, boosting trains classifiers in sequence. Depending on the classification output to the input data weights are assigned giving higher importance to misclassified data points. Afterward also all classifiers give a majority vote. Iterations and a shrinkage parameter need to be specified as parameters.

9Using the programing language R Version: 3.3.2 (R Core Team, 2016) we apply the following R packages: glmnet for the regularized logistic regression (Friedman et al., 2010); for SVM we used e1071 created by Meyer et al. (2015); for random forest we used the randomForest package of Liaw and Wiener (2002); and in the case of ada Boost we use ada (Culp et al., 2016).
3.6 The sample of doctoral students in applied physics and electrical engineering

Various constraints and restrictions of the general approach discussed above had to be made to adapt it to the requirements of the present study. First and most importantly, we require a test sample allowing us to assess the reliability of our co-publication-based approach. To this purpose we utilize the sample of verified student-advisor matches developed by Buenstorf and Geissler (2014), which encompasses the universe of German laser-related doctoral dissertations completed since 1960 and is based on information obtained by university departments as well as the names of examiners provided in the dissertations. In addition to completeness, a further advantage of this dataset is its longitudinal character. However, for our purposes it has one disadvantage. While our algorithm aims at identifying the „true“ advisor, the training set often includes only the first examiner of the dissertation. These two roles may differ, particularly if a doctoral student is advised by a postdoc or a researcher working at a non-university public research organization. The training dataset includes 6,389 doctoral students with dissertations completed from 1960 to 2007. However sample size is reduced for two reasons. First we are stricter about excluding medical dissertation projects than the original authors, and second the doctoral student needs to have at least one publication. This reduces our sample to 3,583 observations, of which 1,686 (47%) have publications identified and 1,367 have their advisor among their co-authors. Only the latter are included in the training dataset.

Second, to ensure that our matching is not biased by field-specific differences in the role of student and advisor characteristics we focus on the broader fields of applied physics and electrical engineering into which much of which laser research falls as a subset. This field is also characterized by a prevalence of publications in WoS-listed journals, a precondition for our approach to work. Specifically, we limit the sample to dissertations classified as science or engineering by the DNB.\footnote{We used the classification scheme of the DNB corresponding to the DDC (Dewey-Decimal Classification) which is an international standard used by libraries for subject classifications.} This entails that we exclude all theses classified as medical dissertations, which account for about 50% of all doctoral theses in Germany. These exclusions reduce the number of authors of doctoral dissertations to 185,860, which also helps to limit the problem of false positive matches caused by homonyms, i.e. non-identical doctoral students and article authors sharing the same names.

Publication data for this set of authors is retrieved from the subset of WoS articles contained in the Science Citation Index dataset. Starting from the full set of all articles, proceedings and reviews with at least one author affiliation in Germany, we further restrict the dataset to articles published in relevant fields. These are identified as follows. First all articles containing „laser“ in the title are selected. The set of relevant articles is then extended using title terms that frequently co-occur with „laser“. The journals in which these articles were published constitute our set of relevant journals.\footnote{We use only WoS publications listed in the science citation index with at least one Ger-}
334,945 articles is covered in the WoS.

The matching procedure of the dissertation data (DNB-dataset) and the publication records (WoS-dataset) resulted in 30,969 true positive matched thesis publication record pairs. We identified 25,856 linked records by finding similarities between the titles of the thesis and the corresponding papers and 22,777 were identified by using matched unique names on both datasets. The doctoral students in the dataset published on average 5.39 papers (median: 3) with 8.46 co-authors (median: 5). Because the sample size of the training data is limited, we use threefold cross validation within the training sample to find best parameter setting (Witten et al., 2011). In the cross-validation procedure we split into test and training set in 1:3 portions, rotating estimation and average results. From the cross-validation we take the best fitting parameter setting and use the algorithm on the evaluation set. As already stated, the models’ predictive power is then evaluated by using a separated partition of the dataset. Evaluation results are presented in Table 1.

Table 1: Training results of advisor classification (selection of best models)

<table>
<thead>
<tr>
<th>Model</th>
<th>+1 (best parameter)</th>
<th>+1 (min recall 0.6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision Recall F1</td>
<td>Precision Recall F1</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.67 0.87 0.77</td>
<td>0.80 0.62 0.71</td>
</tr>
<tr>
<td>SVM</td>
<td>0.80 0.82 0.81</td>
<td>0.90 0.66 0.78</td>
</tr>
<tr>
<td>random Forests</td>
<td>0.84 0.80 0.82</td>
<td>0.84 0.80 0.82</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.80 0.79 0.80</td>
<td>0.80 0.81 0.80</td>
</tr>
</tbody>
</table>

Random Forests outperform the other algorithms in terms of precision and recall. Overall we can obtain an 83% precision and recall rate, which is similar to matching results in other contexts should provide sufficient statistical power to test the above stated hypotheses. We choose to use the random forest model for predicting advisors for the whole sample. Therefore we use the whole test training dataset and apply it to the whole database. Additionally we run algorithms with accepting lower recall aiming at achieving higher precision. We also tested a second set demanding a minimum recall of 0.6 and tried to maximize precision. However, this did not yield a substantial improvement. Nevertheless, we tested all our models applying SVM for classification.

4 Econometric Analysis

Based on the approach detailed in Section 3 we were able to match a total of 25,735 student-advisor pairs in German applied physics and electrical engineer-

---

1269

---

now we select all Web of Science disciplines with a high coverage of relevant journals: „Physics, Applied“; „Optics“; „Physics, Atomic, Molecular - Chemical“; „Engineering, Electrical Electronics“; „Chemistry, Physical“; „Physics, Condensed Matter“; „Materials Science, Multidisciplinary“; „Physics, Multidisciplinary“. Papers with more than 25 authors are not considered.

12Details of the tested parameter settings are provided upon request.
ing for the time period 1975 to 2005. We use this sample to test the hypotheses derived above about determinants of individual scientific fertility (i.e., the likelihood that a given doctoral student becomes a doctoral advisor him- or herself). Specifically, we estimate a set of logit models of the individual likelihood to be identified as an advisor of at least dissertation in our dataset. This likelihood is obviously affected by the time that a given individual defended her own dissertation, which generates a truncation problem that might bias our results. To limit the potential impact of truncation, we further restrict our sample in that we only consider individuals as potential advisors who completed their own dissertation no later than 1995. Accordingly, the final sample is further reduced to 5,373 doctoral students. Descriptive statistics of the used dataset are reported in Appendix B.

Doctoral students are not randomly matched to their advisors. In general, assortative matching of students and advisors can be expected (Azoulay et al., 2017). Students are often deliberately hired from within their future advisors’ networks (Baruffaldi et al., 2016). We are limited in the extent to which we can control for the bias this introduces in the analysis of advisor effects on students. However, we conjecture that the inherent quality of students, as well as the quality of the match with the doctoral advisor, is reflected by the early success of the student. This can be observed in the number of publications a student produces before completing their doctoral degree. Indeed, as is indicated in Table 6, this variable is strongly correlated with advisor characteristics. We further control for the number of co-authors a student accumulates in their early work, which can also be assumed to reflect the quality of students and their matching with the advisor.

Given substantial correlation between some of the explanatory variables, we first use them separately in individual models and then estimate a full model specification including all variables (Models 1 to 8 in Table 2). All models include a set of time dummies (5-year periods) for graduation years and advisors’ first publications, as well as fixed effects for degree-granting universities. We then repeat the same procedure controlling for student quality (Table 3), and again including advisor fixed effects (Table 4, where the number of observations is reduced to 5,140 because all advisors with only a single dissertation in the dataset and with all-positive outcomes are dropped from the estimation).

The role of advisor embeddedness in the scientific community is captured by three different variables: their presence in the principal component of the publication network ($adv_{maincomp}$), their Bonacich centrality ($adv_{bonacich}$) as well as their degree ($adv_{deg}$). Bonacich centrality is proposed by Ballester et al. (2006) as a suitable measure to identify key actors in networks. It goes back to Bonacich (1987) who used it to assess the power of central players, i.e., those who exert strong influence on surrounding players, by counting all paths to other nodes weighted by decay factor increasing with path length. We use a variation of this measure by only considering other advisors as relevant.\footnote{We also further introduce a cut-off value at a path length higher than 5 for computational reasons, as decay factor we chose $2^{- (k-1)}$ depending on the path length $k$.}
In Table 2 we find support for a positive effect of all three variables on students' likelihood to become advisors themselves (their academic „fertility“, as it were) in Models 1 to 3. In the full model, only advisor presence in the principal component remains statistically significant. Controlling for student quality (Table 3) moreover indicates the importance of selection. While we find that better students (in terms of more early publications; cf. Balsmeier and Pellens (2014)) are more likely to become advisors later on\(^{14}\), the coefficient of membership in the principal component is reduced substantially, and both Bonacich centrality and degree lose all of their explanatory power. In the within-estimator framework of the fixed effects models of Table 4, the role of advisor network membership is further qualified, mostly reflecting imprecise measurement because the respective variables mostly vary in the cross section but less so over time. Accordingly, we find only mixed support for the prediction of Hypothesis 1.

Hypotheses 3 and 4 focus on lifecycle effects, i.e. the question whether experienced/older advisors are more likely to turn out fertile students than less experienced/younger ones. We try to capture lifecycle effects by two variables: the number of dissertations advised prior to that of the focal student (\textit{adv\_prestud}), as well as advisor career age at the time of submission (time lag to their first publication; \textit{adv\_careerage}). Our findings point to negative effects of seniority on student fertility, as the coefficients obtained for both variables are mostly negative and often significantly different from zero. In particular, we find strong negative associations between the seniority measures and students’ likelihood to become advisors in the fixed-effects framework of Table 4, which controls for (time-invariant) advisor characteristics. Further support for this interpretation is provided by the significantly negative coefficient of the number of previous citations in the fixed effects model (Model 4 in Table 4).

Finally we look at the peer effects that were subject of Hypotheses 5a and 5b (\textit{fert\_siblings}). A complex pattern emerges. On the one hand, we find robust evidence suggesting a positive peer impact in Models 7 and 8 of Tables 2 and 3. This is consistent with benefits from working in groups where other students successfully embark on academic careers. On the other hand, in line with our above conjecture that peer quality is not independent of advisor quality, positive peer impact appears to be mostly cross-sectional. In contrast, in the fixed-effects models of Table 4, we obtain an even more pronounced negative association between the fertility of contemporaneous „siblings“ and the focal student’s likelihood to become an advisor herself. This indicates that the competitive effect predicted by Hypothesis 5b dominates.

\(^{14}\)The degree (number of co-authors) of students is generally negative and in some models marginally significant. This suggests that conditional on the number of early publications, having more co-authors is associated with a reduced likelihood to remain active in academia and become an advisor. In models without the publication variable, student degree is positive.
Table 2: Logit estimations on determinants of individual scientific fertility

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>adv_maincomp</td>
<td>0.764***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.552**</td>
</tr>
<tr>
<td></td>
<td>(0.212)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.217)</td>
</tr>
<tr>
<td>adv_bonacich</td>
<td></td>
<td>0.009**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.004)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.006)</td>
</tr>
<tr>
<td>adv_deg</td>
<td></td>
<td></td>
<td>0.014***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.004)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.006)</td>
</tr>
<tr>
<td>adv_cit</td>
<td></td>
<td></td>
<td></td>
<td>0.001***</td>
<td></td>
<td></td>
<td></td>
<td>0.0003</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.0001)</td>
<td></td>
<td></td>
<td></td>
<td>(0.0002)</td>
</tr>
<tr>
<td>adv_prevstud</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.015</td>
<td></td>
<td></td>
<td>-0.058***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.013)</td>
<td></td>
<td></td>
<td>(0.016)</td>
</tr>
<tr>
<td>adv_career-age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.038</td>
<td></td>
<td></td>
<td>-0.054*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.026)</td>
<td></td>
<td></td>
<td>(0.028)</td>
</tr>
<tr>
<td>fert_sibling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.375***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.047)</td>
</tr>
<tr>
<td></td>
<td>(0.452)</td>
<td>(0.414)</td>
<td>(0.414)</td>
<td>(0.413)</td>
<td>(0.411)</td>
<td>(0.430)</td>
<td>(0.411)</td>
<td>(0.468)</td>
</tr>
<tr>
<td>Observations</td>
<td>5,373</td>
<td>5,373</td>
<td>5,373</td>
<td>5,373</td>
<td>5,373</td>
<td>5,373</td>
<td>5,373</td>
<td>5,373</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-1,136.543</td>
<td>-1,142.392</td>
<td>-1,140.218</td>
<td>-1,136.364</td>
<td>-1,143.823</td>
<td>-1,143.667</td>
<td>-1,115.821</td>
<td>-1,100.389</td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01
Estimates from logit regression model, (robust) standard errors in parentheses.
All models include a set of time dummies (5-year periods) for graduation years and advisors’ first publications, as well as fixed effects for degree-granting universities.
Table 3: Logit estimations on determinants of individual scientific fertility, including student variables

<table>
<thead>
<tr>
<th>Dependent variable: stud_fert</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
</tr>
<tr>
<td>adv_maincomp</td>
</tr>
<tr>
<td>adv_bonacich</td>
</tr>
<tr>
<td>adv_deg</td>
</tr>
<tr>
<td>adv_cit</td>
</tr>
<tr>
<td>adv_prevstud</td>
</tr>
<tr>
<td>adv_career-age</td>
</tr>
<tr>
<td>fert_sibling</td>
</tr>
<tr>
<td>stud_pub</td>
</tr>
<tr>
<td>stud_deg</td>
</tr>
</tbody>
</table>

| Observations | 5,373 | 5,373 | 5,373 | 5,373 | 5,373 | 5,373 | 5,373 | 5,373 |
| Log Likelihood | -1,056.873 | -1,057.631 | -1,059.088 | -1,058.598 | -1,058.054 | -1,057.505 | -1,048.389 | -1,037.566 |

Note: *p<0.1; **p<0.05; ***p<0.01
Estimates from logit regression model, (robust) standard errors in parentheses.
All models include a set of time dummies (5-year periods) for graduation years and advisors’ first publications, as well as fixed effects for degree-granting universities.
Table 4: Fixed-effect Logit estimations on determinants of individual scientific fertility, including student variables

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>stud_fert</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>adv_maincomp</td>
<td>0.366</td>
</tr>
<tr>
<td></td>
<td>(0.434)</td>
</tr>
<tr>
<td>adv_bonacich</td>
<td>−0.031***</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
</tr>
<tr>
<td>adv_deg</td>
<td>−0.016</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
</tr>
<tr>
<td>adv_cit</td>
<td>−0.001**</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
</tr>
<tr>
<td>adv_prevstud</td>
<td>−0.058**</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
</tr>
<tr>
<td>adv_career-age</td>
<td>−0.151***</td>
</tr>
<tr>
<td></td>
<td>(0.056)</td>
</tr>
<tr>
<td>fert_sibling</td>
<td>−1.003***</td>
</tr>
<tr>
<td></td>
<td>(0.117)</td>
</tr>
<tr>
<td>stud_pub</td>
<td>0.355***</td>
</tr>
<tr>
<td></td>
<td>(0.035)</td>
</tr>
<tr>
<td>stud_deg</td>
<td>−0.044*</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
</tr>
<tr>
<td>Constant</td>
<td>−19.952</td>
</tr>
<tr>
<td></td>
<td>(16,877.470)</td>
</tr>
<tr>
<td>Observations</td>
<td>5,140</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>−668.809</td>
</tr>
<tr>
<td>Akaike Inf. Crit.</td>
<td>3,427.619</td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01
Estimates from fixed-effect logit regression model, standard errors in parentheses.
All models include a set of time dummies (5-year periods) for graduation years.
5 Discussion and conclusions

In this paper we presented a novel approach to identify the advisors of doctoral students based on co-publications and machine learning of matching algorithms. This approach enabled us to construct and analyze a large-scale dataset of German doctoral graduates in applied physics and electrical engineering from 1975 to 2005. We analyzed factors associated with the subsequent likelihood of these individuals to become doctoral advisors themselves, i.e. to be „fertile“ as academics and play an active role in the self-reproduction of science. We obtained substantial evidence indicating that advisor seniority is associated with a decreased likelihood of students to become advisors themselves. This finding, which was robust to controlling for student quality and advisor heterogeneity, suggests that finding good advisor early in their career is helpful for the academic career of doctoral students. One interpretation is that younger advisors are more in touch with current developments in their fields, and that this is more important than their academic reputation based on earlier success or their position in the publication network. This interpretation is consistent with the limited importance that we found for advisors’ network embeddedness or citations to their prior work. In addition, our results suggest a relevant role of competition for academic career opportunities among the students co-advised by an advisor at the same time. The dynamics that our results imply are consistent with dynastic relationships in science. However, in contrast to what is argued by Horta et al. (2010), they seem to be more suggestive of meritocratic processes than of navel gazing.

Both the methodological approach to match students and advisors and the specific results on academic „fertility“ are not without limitations. First and perhaps most importantly, we can only match individuals for whom we can co-authored publications in openly accessible databases. This excludes fields and disciplines in which the respective type of publication is uncommon. This holds for large parts of the social sciences and humanities. Second, our approach identifies those individuals as advisors who leave their mark on students in terms of how they conduct research. While we think that these individuals are indeed the key persons in socializing the next generations of scientists, they may differ from those who are officially listed as advisors, e.g., university professors „filling in“ as examiners for researchers lacking the formal qualification to do so. Third, we have so far only considered student-advisor pairs where both parties completed their dissertations at a German university. In an increasingly globalized system, this is an increasingly limiting restriction. It is likewise not clear whether our above findings on the determinants of student „fertility“ are generalizable to other fields and countries. Moreover, while we tried to control for the quality of students and the student-advisor match, a causal interpretation of our results would require randomized matching which is not what we have in our data.
References


### A Data quality proofs

#### A.1 Web of Science Author Name Disambiguation

Testing the assignment quality of the WoS disambiguation procedure we compare the number of publications attributed to very frequent and rare names.
To get accurate numbers of the name frequencies we use the DNB database on doctoral students. Because foreign names are overrepresented in unique name combinations we consider only doctoral students with German nationality in the two name groups. Doctoral students from abroad might suffer from systematic lower publication records, since they are more likely to leave Germany after graduation. We select 100 most frequent name combinations (surname and first name initial) in the whole DNB dataset (which covers about one million records) and compare these against names which occur only once. The distribution of the publication oeuvres should be equal if our disambiguation procedure applies. Figure 2 compares the two distributions, which indeed appear to be equal. We cannot find a significant difference between the mean number of publications authored by researchers with one of the 100 most frequent names and the ones with very unique names (means: 12.37, 13.36; p-value: 0.72). The median equals 3 for both groups.

![Graph showing density of publications assigned to 100 most frequent (German) names and unique (German) names.]

**Figure 2:** Density of publications assigned to 100 most frequent (German) names and unique (German) names

### A.2 Publication Dissertation Record Linkage

As robustness check of the matching between the dissertation data with the WoS scientists we use the first names, which we could not use in large scale. We
find for 6,466 pairs the full first name given on both sources. By comparing all first names manually finding 257 pairs with different name. This results in an error rate of 3.97% (or 96% correctly matched pairs).

B Descriptive statistics

Table 5: Summary Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>stud_fert</td>
<td>0.06</td>
<td>0.23</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>stud_pub</td>
<td>2.81</td>
<td>2.62</td>
<td>1</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>stud_deg</td>
<td>3.38</td>
<td>3.71</td>
<td>0</td>
<td>2</td>
<td>43</td>
</tr>
<tr>
<td>adv_maincomp</td>
<td>0.82</td>
<td>0.39</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>adv_bonacich</td>
<td>11.18</td>
<td>16.15</td>
<td>0.00</td>
<td>3.19</td>
<td>84.62</td>
</tr>
<tr>
<td>adv_deg</td>
<td>11.18</td>
<td>11.45</td>
<td>1</td>
<td>8</td>
<td>212</td>
</tr>
<tr>
<td>adv_cit</td>
<td>224.37</td>
<td>305.15</td>
<td>0</td>
<td>117</td>
<td>2,654</td>
</tr>
<tr>
<td>adv_prev-stud</td>
<td>6.04</td>
<td>6.09</td>
<td>1</td>
<td>4</td>
<td>40</td>
</tr>
<tr>
<td>adv_career-age</td>
<td>12.45</td>
<td>4.92</td>
<td>0</td>
<td>13</td>
<td>21</td>
</tr>
<tr>
<td>fert_siblings</td>
<td>0.65</td>
<td>1.05</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 6: Correlations

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
<th>(10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>stud_fert</td>
<td>1</td>
<td>0.21</td>
<td>0.08</td>
<td>0.05</td>
<td>0.02</td>
<td>0.04</td>
<td>0.07</td>
<td>-0.02</td>
<td>-0.02</td>
<td>0.12</td>
</tr>
<tr>
<td>stud_pub</td>
<td>0.21</td>
<td>1</td>
<td>0.52</td>
<td>0.20</td>
<td>0.33</td>
<td>0.34</td>
<td>0.34</td>
<td>0.09</td>
<td>0.17</td>
<td>0.28</td>
</tr>
<tr>
<td>stud_deg</td>
<td>0.08</td>
<td>0.52</td>
<td>1</td>
<td>0.16</td>
<td>0.34</td>
<td>0.57</td>
<td>0.23</td>
<td>0.13</td>
<td>0.16</td>
<td>0.23</td>
</tr>
<tr>
<td>adv_maincomp</td>
<td>0.05</td>
<td>0.20</td>
<td>0.16</td>
<td>1</td>
<td>0.33</td>
<td>0.26</td>
<td>0.27</td>
<td>0.19</td>
<td>0.30</td>
<td>0.24</td>
</tr>
<tr>
<td>adv_bonacich</td>
<td>0.02</td>
<td>0.33</td>
<td>0.34</td>
<td>0.33</td>
<td>1</td>
<td>0.61</td>
<td>0.60</td>
<td>0.36</td>
<td>0.43</td>
<td>0.35</td>
</tr>
<tr>
<td>adv_deg</td>
<td>0.04</td>
<td>0.34</td>
<td>0.57</td>
<td>0.26</td>
<td>0.61</td>
<td>1</td>
<td>0.50</td>
<td>0.42</td>
<td>0.32</td>
<td>0.42</td>
</tr>
<tr>
<td>adv_cit</td>
<td>0.07</td>
<td>0.34</td>
<td>0.23</td>
<td>0.27</td>
<td>0.60</td>
<td>0.50</td>
<td>1</td>
<td>0.37</td>
<td>0.28</td>
<td>0.50</td>
</tr>
<tr>
<td>adv_prev-stud</td>
<td>-0.02</td>
<td>0.09</td>
<td>0.13</td>
<td>0.19</td>
<td>0.36</td>
<td>0.42</td>
<td>0.37</td>
<td>1</td>
<td>0.47</td>
<td>0.39</td>
</tr>
<tr>
<td>adv_career-age</td>
<td>-0.02</td>
<td>0.17</td>
<td>0.16</td>
<td>0.30</td>
<td>0.43</td>
<td>0.32</td>
<td>0.28</td>
<td>0.47</td>
<td>1</td>
<td>0.20</td>
</tr>
<tr>
<td>fert_siblings</td>
<td>0.12</td>
<td>0.28</td>
<td>0.23</td>
<td>0.24</td>
<td>0.35</td>
<td>0.42</td>
<td>0.50</td>
<td>0.39</td>
<td>0.20</td>
<td>1</td>
</tr>
</tbody>
</table>
Are Great Researchers Terrible Teachers?
How research and teaching performance relate at U.S. universities

Dakota Murray¹, Huimeng Zhao¹, Vanessa Minik¹, Nicolas Bérubé², Vincent Larivièrè², and Cassidy R. Sugimoto¹

¹dakmurrr@iu.edu; huimzhao@iu.edu; vminik@iu.edu; sugimoto@indiana.edu
Indiana University—Bloomington (U.S.A.)
²nicolas.berube.3@umontreal.ca; vincent.lariviere@umontreal.ca
Université de Montréal (Canada)

Abstract
University professors engage in a variety of tasks in the domains of research, teaching, and service. There exists an abundance of research exploring a professor’s performance in any one of these domains, yet few have explored how they affect each other. To explore the relationship between a scholar’s research and teaching performance, we merged a dataset provided by the company Academic Analytics, which provides research performance data to U.S. universities, with another dataset collected from the popular teacher-evaluation website Rate My Professor. Using these data, we find that, controlling for all other variables, gender, discipline, age, and class difficulty were strong and significant predictors of professor rating. Differences are found by gender and by discipline, with women with the highest teaching evaluations tending to have the lowest research performance. The average quality scores of men are higher across all the broad disciplinary areas, with a wide gap in Engineering but the smallest in the Medical Sciences. Male professors are also rated as having more difficult classes, except in the humanities. This research is only the beginning of what large-scale quantitative analysis promises to contribute to the understanding of research and teaching performance.

Conference Topic
Science Policy; Research Evaluation; Teaching Assessment

Introduction
Criteria for promotion, tenure, and merit review revolve around three main tasks: research, teaching, and service. However, these categories are not weighted equally: research performance is the most highly weighted factor, followed by teaching and service (Boyer, 1990). It has been suggested that this is a zero-sum game—in which emphasis on one aspect of the professorial life comes at the expense of others. This so-called scarcity model of teaching and research had found evidence of the divergence between teaching and research activities (Lesyte, Enders, & De Boer, 2009). Other scholars have found correlations between research and teaching, demonstrating that active researchers tend to also be good teachers (Halse, Hobson, & Jones, 2007). A third body of research rejects the assumption of any generalizable relationship between teaching and research (Hattie & March, 1996; Figlio & Schapiro, 2017). Despite these exploratory studies, there is no consensus on the relationship between research and teaching in academic institutions. The extant literature suffers from small sample sizes, limited disciplinary diversity, a focus on tenured professors, and other factors that limit the external validity and generalizability of the findings. The present study seeks to address this gap by providing a large-scale study of the relationship between teaching and research at academic institutions in the United States.

There is considerable research on the relationship between demographic characteristics of the faculty, their research, and their teaching performance. Women, for example, tend to be less productive, hold less funding, and receive fewer citations (Larivièrè, Ni, Gingras, Cronin, & Sugimoto 2013; Larivièrè, Vignola-Gagnè, Villeneuve, Gélinas, & Gingras 2011) than their male counterparts. Teaching evaluations can be biased by the gender of both the professor and the student, with women receiving higher evaluations from female
students and lower ratings from male students (Basow, 1995). External factors, such as discipline (Cramer & Alexitch, 2000) and class difficulty can also affect student evaluations. For example, students in the Social Sciences and Humanities tend to rate their professors higher than students in the Sciences (Cramer & Alexitch, 2000). Several studies have examined correlations between the teaching evaluations and professor characteristics. For example, professors belonging to racial minorities are evaluated more negatively than white faculty on the site, and were also rated as holding easier courses (Reid, 2010). Other factors, such as age and attractiveness (Sohr-Preston, Boswell, McCaleb, & Robertson, 2016) as well as industry interactions (Gulbrandsen & Smeby, 2005) also influence the ratings received.

But data on teaching evaluations is limited, since these indicators tend to be held privately by institutions and instructors. One exception is the Rate My Professor (RMP) website, which provides a public mechanism for reporting and analysing teaching evaluations. This site, which allows student to review their instructors anonymously, was released in 1999 and has seen wide use in the United States.

This creates a tangled mess of variables, obscuring the relationship between teaching and research performance. To address this, we analyse the relationship between teaching and research portfolios of nearly 40,000 unique professors in the United States, across the disciplinary spectrum and taking into account demographic information, such as age and gender. This allows the first large-scale exploration of the degree to which research indicators are predictive of teaching evaluations, accounting for differences in the professor’s institution, discipline, age, and gender.

Data and Methods

Four sources of data are aggregated for this study, of which the two most important are 1) the 2014 release of data provided by contract from Academic Analytics (AA2014) and 2) professor ratings and comments sourced from the popular website Rate My Professor. Academic Analytics is a proprietary company that has constructed a database, using public or voluntarily supplied data, of individual professors and publications affiliated with one of more than 385 U.S. universities, and sells access and analysis of this data to university administrators for benchmarking purposes. Rate My Professor (RMP) is a website, popular among college students, that crowd-sources teaching evaluations, allowing (presumed) students to leave comments and rate their teachers along dimensions of helpfulness, clarity, course difficulty, and overall quality.

AA2014 contains information about individual professor’s scholarly performance and research funding between 2010 and 2014, except for number of books, which goes back to 2005, and for awards, which have an indefinite timeline. There are 301,673 total records in AA2014; however, records are duplicated for every department in which a professor holds an affiliation, and as such there are only 163,892 unique individuals represented.

We created a script to aggregate all of the individuals and comments on the Rate My Professor website into a single dataset (RMP). This dataset was collected in early September 2016, and includes profiles of 878,748 individuals, each manually entered by users of the sites, teaching in higher education institutions of all kinds in the U.S., U.K., and Canada. For each rating, a user is asked to rate the professor’s course on various dimensions, leave comments, and add community tags to summarize their experience. At the date of collection, there were a total of 14,735,325 user ratings.

These datasets were merged in order to evaluate the relationship between scientific and teaching indicators. Using the first name, last name, institution, and department of each professor, we matched 50,172 unique professors common to both AA2014 and RMP, comprising 88,980 total records when considering multiple departmental affiliations. Many professors appear in the RMP dataset, but have no ratings, thus leaving 37,202 individuals
after including only those with least one rating. To be consistent with AA2014, we selected only those RMP ratings occurring after 2010 (i.e., 531,359 total ratings) and using these, calculated the mean overall quality, helpfulness, clarity, and easiness for every individual. We also matched institutional names with data from the US News and World Report (USNEWS) and Carnegie Classification of Institutions of Higher Education (CC) datasets, enriching our data. Finally, we superimposed our own simple disciplinary taxonomy over Academic Analytics’ more fine-grained taxonomy. Details about each of the variables in the final aggregated dataset can be found in Table1.

Table 1: Field descriptions for Academic Analytics and Rate My Professor datasets

<table>
<thead>
<tr>
<th>Academic Analytics</th>
<th>Rate My Professor</th>
</tr>
</thead>
<tbody>
<tr>
<td>DegreeYear</td>
<td>Overall Quality</td>
</tr>
<tr>
<td>Gender</td>
<td>Easiness</td>
</tr>
<tr>
<td>Level4Name</td>
<td>Helpfulness</td>
</tr>
<tr>
<td>ArticleCount</td>
<td>Clarity</td>
</tr>
<tr>
<td>CitationCount</td>
<td></td>
</tr>
<tr>
<td>GrantCount</td>
<td></td>
</tr>
<tr>
<td>AwardCount</td>
<td></td>
</tr>
<tr>
<td>ConfprocCount</td>
<td></td>
</tr>
<tr>
<td>Bookcount</td>
<td></td>
</tr>
<tr>
<td>Overall Quality</td>
<td></td>
</tr>
<tr>
<td>Acceptance Rate</td>
<td></td>
</tr>
<tr>
<td>docresflag</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DegreeYear</td>
<td>Year that researcher obtained their degree</td>
</tr>
<tr>
<td>Gender</td>
<td>Gender of individual, assigned either manually, or with a name-based algorithm gender-assignment algorithm by Academic Analytics, limited to male, female, or unknown</td>
</tr>
<tr>
<td>Level4Name</td>
<td>An invention of ours, this variable is an abstraction of Academic Analytics’ discipline classification structure into one more commonly understood by general researchers.</td>
</tr>
<tr>
<td>ArticleCount</td>
<td>Count of peer-reviewed journal articles published in 2011, 2012, 2013 and 2014; data is derived from CrossRef, as well as data self-reported by universities working with Academic Analytics. For co-authored articles, all authors are credited.</td>
</tr>
<tr>
<td>CitationCount</td>
<td>Citations to articles and proceedings from 2010, 2011, 2012, 2013 and 2014, derived from the CrossRef citation-linking network. Self-citations included.</td>
</tr>
<tr>
<td>GrantCount</td>
<td>Count of grants data from 13 federal agencies and two non-federal sources matched to principal investigators. For NIH, NSF and NOAA grants, matching includes co-principal/multi-principal investigators</td>
</tr>
<tr>
<td>AwardCount</td>
<td>The count of awards among honorific awards from 821 governing societies that are open to all people in a discipline, sub discipline, or a large subset of people at a national or international level, and that can be matched to</td>
</tr>
<tr>
<td>ConfprocCount</td>
<td>Count of conference proceedings published in 2011, 2012, 2013 and 2014, as well as data self-reported by universities working with Academic Analytics. All co-authors are credited with writing the article.</td>
</tr>
<tr>
<td>Bookcount</td>
<td>Count of time appearing as author, co-author, editor, co-editor, and translator of books published in 2005-2014 (inclusive), and appearing in Baker &amp; Taylor, The British Library, and internal collection efforts of Academic Analytics. Introductions, forewords, afterwards, and citations are not included in the 2014 Academic Analytics dataset. Due to their limited use of DOIs, chapters are limited in the 2014 release.</td>
</tr>
<tr>
<td>Overall Quality</td>
<td>The average of a professor's Helpfulness and Clarity ratings, intended as a general measure of professor quality.</td>
</tr>
<tr>
<td>Easiness</td>
<td>A measure of the perceived easiness of the class. Rated on an integer scale between one and five, where one is considered hard, and five is considered easy.</td>
</tr>
<tr>
<td>Helpfulness</td>
<td>A measure of the perceived helpfulness of the professor, such as their approachability and willingness to assist students. Rated on an integer scale from one to five, where one is considered useless, while five is considered very helpful</td>
</tr>
<tr>
<td>Clarity</td>
<td>A measure of a professor’s perceived clarity, such as their communication style and ability to clearly present course material. Rated on an integer scale between one and five, where one is considered confusing, and five is considered clear.</td>
</tr>
<tr>
<td>Acceptance Rate</td>
<td>The undergraduate acceptance rate of a university, as published by US News and World Report.</td>
</tr>
<tr>
<td>docresflag</td>
<td>Indicates whether the institution conferred research/scholarship doctoral degrees during the timespan of data collection</td>
</tr>
</tbody>
</table>
Limitations

Concerns have been raised about the validity of both RMP and AA2014. It has been shown that the ratings on RMP may differ from formal in-class evaluations (Coladarci & Kornfield, 2007), and that ratings may lack external validity, with students rating a course on their first day, or years after (Davison & Price 2009). There are also entirely fabricated records; for example, RMP includes individuals such as Albus Dumbledore, professor of transfiguration at Hogwarts School of Witchcraft and Wizardry, with 121 student ratings. Despite these issues, there is also research that supports the validity of evaluations on the website (Sonntaga, Bassett & Snyder 2009), and RMP remains the only large-scale publicly available dataset for teaching evaluations. Similarly, the validity of AA2014 has been called into question. However, given the proprietary nature of the data, the allegations have been largely anecdotal. A large-scale validation exercise remains necessary to properly identify the degree of accuracy in both datasets. We acknowledge this as a limitation of the present analysis.

Results and Discussion

A regression was conducted with the mean overall quality rating from RMP as the dependent variable. Controlling for all other variables, research indicators do not demonstrate a strong effect on course evaluation ratings: Conference proceeding count has a small non-significant effect; Article count, while significant at the 0.005 level, has a negligible effect; Citation count have a trivial non-significant positive effect; Grant count has a negligible and non-significant negative effect. Of all the research indicators, having written or contributed to at least one book and having won at least one award have the strongest effects, both positive, yet this effect is still minor and non-significant. These findings reinforce the notion of a non-linear relationship between research and teaching indicators.

Table 2: Estimate of fixed effects with mean rating on RMP as dependent variable

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t</th>
<th>Sig</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.092</td>
<td>0.16</td>
<td>155285.52</td>
<td>31.746</td>
<td>0.000</td>
<td>4.778 - 5.407</td>
</tr>
<tr>
<td>Gender = Male</td>
<td>0.164</td>
<td>0.0144</td>
<td>25269.322</td>
<td>11.405</td>
<td>0.000</td>
<td>0.136 - 0.191</td>
</tr>
<tr>
<td>Gender = Female</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gender = Known</td>
<td>0.096</td>
<td>0.138</td>
<td>25262.8</td>
<td>0.696</td>
<td>0.486</td>
<td>-0.175 - 0.367</td>
</tr>
<tr>
<td>Gender = Unknown</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[AWARDSWON = 0.0]</td>
<td>-0.0267</td>
<td>0.0125</td>
<td>25260.73</td>
<td>-2.127</td>
<td>0.033</td>
<td>-0.0513 - 0.0021</td>
</tr>
<tr>
<td>[AWARDSWON = 1.0]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Discipline = Engineering</td>
<td>-0.1</td>
<td>0.0228</td>
<td>24471.628</td>
<td>-4.397</td>
<td>0.000</td>
<td>-0.145 - 0.0555</td>
</tr>
<tr>
<td>Discipline = Humanities</td>
<td>0.141</td>
<td>0.0359</td>
<td>25266.1</td>
<td>3.938</td>
<td>0.000</td>
<td>0.071 - 0.212</td>
</tr>
<tr>
<td>Discipline = Medical Sciences</td>
<td>0.0689</td>
<td>0.0228</td>
<td>25107.37</td>
<td>3.023</td>
<td>0.003</td>
<td>0.0242 - 0.114</td>
</tr>
<tr>
<td>Discipline = Natural Sciences</td>
<td>-0.0433</td>
<td>0.02</td>
<td>25148.29</td>
<td>-2.165</td>
<td>0.030</td>
<td>-0.0826 - 0.0041</td>
</tr>
<tr>
<td>Discipline = Social Sciences</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[docresFlag = 0.0]</td>
<td>-0.142</td>
<td>0.182</td>
<td>93.885</td>
<td>-0.783</td>
<td>0.436</td>
<td>-0.504 - 0.219</td>
</tr>
<tr>
<td>[docresFlag = 1.0]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[Has Books = 0.0]</td>
<td>-0.0125</td>
<td>0.0146</td>
<td>25245.8</td>
<td>-0.859</td>
<td>0.391</td>
<td>-0.041 - 0.016</td>
</tr>
<tr>
<td>[Has Books = 1.0]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ArticleCount</td>
<td>-0.001</td>
<td>0.0003</td>
<td>25262.83</td>
<td>-3.317</td>
<td>0.001</td>
<td>-0.0016 - 0.00042</td>
</tr>
<tr>
<td>ConfprocCount</td>
<td>-0.001</td>
<td>0.0008</td>
<td>25269.95</td>
<td>-1.314</td>
<td>0.180</td>
<td>-0.0025 - 0.0005</td>
</tr>
<tr>
<td>Log(CitationCount)</td>
<td>0.0099</td>
<td>0.0044</td>
<td>25153.54</td>
<td>2.252</td>
<td>0.024</td>
<td>0.0013 - 0.0185</td>
</tr>
<tr>
<td>Log(GrantCount)</td>
<td>-0.0078</td>
<td>0.0057</td>
<td>25270.0</td>
<td>-1.372</td>
<td>0.170</td>
<td>-0.019 - 0.0034</td>
</tr>
<tr>
<td>Age</td>
<td>0.012</td>
<td>0.0006</td>
<td>25269.67</td>
<td>19.285</td>
<td>0.000</td>
<td>0.0109 - 0.013</td>
</tr>
<tr>
<td>RMP Course Difficulty</td>
<td>-0.491</td>
<td>0.006</td>
<td>25269.98</td>
<td>-7.569</td>
<td>0.000</td>
<td>-0.504 - 0.478</td>
</tr>
<tr>
<td>USNEWS Acceptance Rate</td>
<td>0.002</td>
<td>0.0006</td>
<td>142.138</td>
<td>-3.456</td>
<td>0.001</td>
<td>-0.0033 - 0.0009</td>
</tr>
</tbody>
</table>

a. Dependent Variable: RMP Overall Quality
b. This parameter is set to zero because it is redundant
However, even when controlling for all other variables, the regression did provide evidence that gender, discipline, age, and class difficulty have the strongest effects and are significant predictors of professor rating \( (p < 0.000) \). Course difficulty, followed by gender, has the strongest effect of any variable. Men tend to receive higher ratings than women, and as course difficulty increases, average ratings tend to decrease. Analysing interaction effects among these variables may provide additional insight into the relationship between research and teaching variables. For example, Figure 1 depicts the gendered differences in each of the research indicators, by overall RMP score. As shown, women with the highest teaching evaluations tend to be those with the lowest article and citation counts. However, the inverse is true for male professors.

![Figure 1. Scholarly performance indicators by gender and average overall teaching quality. Brackets signify inclusivity of ratings, and parentheses indicate exclusivity](image)

Disciplinary differences by gender are also apparent, showing large statistically significant effects on teaching ratings. Although the average quality scores for men are higher across all the broad disciplinary areas, the gap is much smaller in the Medical Sciences and highest in Engineering. Furthermore, male professors are also rated as having more difficult courses, on average, with the exception of the Humanities—where female professors receive a higher average rating.
Conclusion
This paper begins to explore the various relationships between teaching and research, taking into account characteristics of the professor and institution. The regression suggests a non-linear relationship between research and teaching indicators. It also reveals that gender, discipline, age of professor, and course difficulty strong predictors of teaching evaluation ratings. Further analysis is necessary to more fully understand the complicated relationship when examining teaching and research indicators by gender and discipline. This aggregated dataset provides a useful and large-scale approach to begin analysing these questions.

Acknowledgments
This work is funded by the National Science Foundation award #1561299 (EAGER: Illuminating the role of science funding on disparities in science).

References


Research or management? An investigation of the impact of administrative roles on the research performance of academic administrators

Yuehua Zhao¹ Wen Lou²* Yuchen Chen²
¹ yuehua@uwm.edu University of Wisconsin-Milwaukee (United States)
² wlou@infor.ecnu.edu.cn, 893781834@qq.com East China Normal University (China)

Abstract
The present study aims to investigate whether the research performance of academic administrators has been affected by their administrative services in a quantitative way. We sampled 116 academic administrators from 26 universities. Statistical methods were applied to compare the research performance of the sampled administrators before and after they accepted the administrative roles. The results suggest that administrative services have strongly affected academic leaders’ research performance. The extent of such impact varies in different disciplines. The impact appears to be higher on administrators serving in higher-ranking universities. In addition, the research outputs of both university presidents and department deans have been substantially influenced by the administrative services, whereas the impact on the university presidents is slightly stronger than on the department deans.

Conference Topic
The application of informetrics on evaluation; Science policy and research assessment

Introduction
It has become more and more common worldwide that after years of excellent teaching and research services in a university, a faculty would serve as a department or school leaders (Carroll, 1991). Besides the administrative affairs based on the establishment, monitoring, and modification of programs and curricula, staff matters, and student affairs (Boyko & Jones, 2010), the leaders are involved with scientific research and represent the top research level of schools and departments. The time management to balance administrative roles and research has caught the attention of higher education managers (Magana & Niebel, 1976). Investigation on job satisfaction in universities claimed that deans and chairs reported feeling overwhelmed at their job every day (Morris & Laipple, 2015). They believed it resulted from the lack of full preparation for the administrative roles because their career paths were only formed by years of scientific research (Amundsen & Martinsen, 2014). Therefore, we wonder whether their research productivity has been affected by administrative roles.

The relationship between research performance and administrative roles of university professors has been discussed in a few studies. Gu et al. (2014) compared the publications of 18 presidents from six universities on the period of both service and retiring and found that presidents in the US had been barely influenced by presidency service in terms of the research performance. Their study is quite close to our aiming but limited to the small objects and inadequate of statistical tests. García-Gallego et al. (2015) conducted a comparison analysis on the impact of research and administrative duties on teaching performance. On the contrary, more studies discussed the influences of university leaderships that can bring to staff management (Sheikha & Younis, 2006; Siddique et al., 2011) and the quality of teaching and learning (Hammond, 1999; Nkanata, 2013).

In summary, there were few empirical studies conducted on research performance of academic leaders. In this study, academic administrators refer to individuals who serve as leadership and
management roles within the university systems worldwide. Our aim is to explore, by answering the following four questions, how the administrative role as a university president or a department dean affects one’s research performance as a researcher.

RQ1. Do administrative services affect academic administrators’ research works?
RQ2. Does the impact of administrative roles on administrators’ research performance differ among different disciplines?
RQ3. Does the impact of administrative roles on administrators’ research performance differ among universities in different ranking levels?
RQ4. Does the impact on administrators’ research performance differ among different administrative roles?

Method

Sampling strategy
At first, 29 universities were selected based on 2017 Best Global Universities Rankings established by the U.S. News & World Report, whose rankings are at 1 to 10, 51 to 60, and 101 to 110. To explore the differences among different disciplines, we attempted to select six types of schools and departments, including art and humanities, psychology, law, business and management, engineering, and medical. We looked into the official website of every university to ensure these departments exist in the university. It appeared that the most common departments are Art, Business and Management, Engineering, and Medical. Besides that, the following two universities were excluded due to the lack of most of the departments. California Institute of Technology only has business and engineering schools, and Mount Sinai School of Medicine only has the medical department. Therefore, 27 universities with their presidents and four types of academics were selected for the next step.

To compare the administrative roles’ influence from different duties, the president of each university and the dean of each school or department were both collected. Firstly, and easily, we collected the name of every university president. Secondly, we looked through, in turn, the faculties, schools, and departments’ leaders, since some universities have mixed administrative structure. For instance, if a university has both Faculty of arts and humanities and School of Arts, we selected the leader of Faculty of arts and humanities. Thirdly, the leadership of schools or departments contains many different titles, such as chairperson, director, dean, research dean, and so on. We selected the most important leaders to guarantee there would be only one leader in each school or department. Interim leaders were selected only if a university or department had no official announced leaders yet. Additionally, University of Munich was excluded since we could not get more information of leaders due to the access of website and language issue. So far, 130 administrators were chosen in total.

Data collection
The very first information that we need to consider is which year these leaders were announced as the administrative roles. We collected the exact year from looking through those leaders’ CVs and LinkedIn pages, and university news and welcome letters on the websites if their profiles did not mention the year. Meantime, their career paths were recorded in the documents including every affiliation they have been as doctoral students or professors, which enabled us to identify their affiliations in Web of Science databases (WOS).

Full records of documents that published by sampled leaders were downloaded from WOS core collection. Data collection were conducted during January 31st to March 27th, 2017. We paid more attention to author addresses and research domains and areas to identify every leader due to the duplicated names and flaws of author identifier. Considering the different publication styles in different disciplines, comments, reviews, and other types of documents were all included in this paper. In the end, 116 leaders with publications will be discussed in the following sections since there were no publications of 14 leaders in WOS.
We measured the publication counts of each sampled leader during the following three periods: the pre-position period, the reference period, and the in-position period. The pre-position period means the years from when a leader published his or her first paper until the year he or she became the leader. The in-position period represents the years that a leader has been holding the position since the year he or she was announced to be the leader until the day we collected data. The reference period is inspired by the idea of Impact Factor. It balances the different periods of in-position and pre-position by limiting the time length of the pre-position period to make the length of time the same as the in-position period. For example, Michael D. Smith has been the dean of Harvard Faculty of Arts and Sciences for ten years since 2007. Then the number of publications during his reference period was counted from 1997 to 2006, which is the same length of time of his occupation. In addition, we excluded the publications published in the year that a faculty member became the administrator since we could not separate publications in this year from either pre-position period or in-position period.

Data analysis

To answer the above research questions, we adopted statistical methods to investigate the impact of administrative roles on administrators’ research performance in terms of the numbers of publications. The following null hypotheses were proposed and tested in the study.

H01: There is no significant difference between administrators’ research performance during the reference period and the in-position period.

H02: There are no significant differences among the four disciplines in terms of the impact of the administrative roles on the performance of academic administrators.

H03: There is no significant difference among the three levels of the university ranking in terms of the impact of the administrative roles on the performance of academic administrators.

H04: There is no significant difference between the two administrative roles in terms of the impact of the administrative roles on the performance of academic administrators.

The parametric tests (e.g. t-tests, analysis of variance) make assumptions about the shape of the population distribution (e.g. normally distributed), while non-parametric techniques do not include assumptions about the underlying population distribution (Pallant, 2013). Therefore, in this study, a series of non-parametric tests (e.g. Wilcoxon signed-rank test, Kruskal-Wallis H test) were performed to examine the null hypotheses. The significance level (α) for all tests is equal to 0.05. If the resultant p-value of a null hypothesis test is smaller than 0.05, the null hypothesis is rejected. Otherwise, the null hypothesis fails to be rejected. In the Kruskal-Wallis H test, the result includes both χ²-value and p-value. The χ²-value is presented as χ²(df, n). Here df stands for the degrees of freedom, and n stands for the sample size. Effect size indicates the influence of the independent variable (Pallant, 2013). In this study, the effect size statistics was reported as the r-value, and the median was reported as the Md value. We adopted the difference between the number of publications during the reference period and the in-position period to measure the impact of administrative services on faculty members’ research performance.

Results and Discussion

An overview

In total, we sampled 116 academic administrators, including 94 department deans and 22 university presidents, from 26 universities. The distribution of the publications in different time spans is displayed in Figure 1 where the X-axis represents each leader, and the Y-axis is the number of papers they published. The three lines, separately, indicate the number of publications yielded during the three periods: the pre-position period, the reference period, and the in-position period. In Figure 1, the pre-position publications outnumbered the in-position publications substantially. Therefore, we also calculated the papers published during the same
length of time they spent in the position before their posts to avoid the bias of time span. Given the consideration of the time spent, the decline of the research outputs was still evident.

Figure 1. Distributions of research performance of administrators during different periods

The Wilcoxon signed-rank test (also referred to the Wilcoxon matched pairs signed ranks test) is designed for the repeated measures (Pallant, 2013), and thus it was used to the to examine the hypothesis \( H01 \). The test results demonstrated that there was a significant difference between administrators’ research performance during the reference period and the in-position period \((z=-8.087, p=0.00<0.05)\). The resultant effect size \( r \) was 0.531, indicating a large effect size using Cohen’s (1988) criteria of 0.1=small effect, 0.3=medium effect, 0.5=large effect (Pallant, 2013). It suggests that administrative services significantly affect academic administrators’ research works.

Before performing the administrative services, on average, each leader yielded 111 publications. The average number of publications dropped down to six during their services. Among the investigated administrators, 114 out of 116 (98.3%) administrators performed better before the posts. On average, people published 105 more papers before they became the administrators. Considering the length of time, there are still 108 leaders produced more papers during the reference period. The average decline appeared to be 18 papers. It showed that most of the faculties were highly productive before accepting the administrator positions, but their research productivities were apparently affected by the administrative services.

Comparison among disciplines

In this study, the dependent variable for the Kruskal-Wallis H test was the difference between the number of publications during the reference period and the in-position period. The independent variable was the four defined disciplines: Art, Business and Management, Engineering, and Medical. The results of the Kruskal-Wallis H test revealed that there were statistically significant differences among the four defined disciplines in terms of the impact of the administrative roles on the research performance of academic administrators \( (\chi^2(3, n = 94) = 18.224, p = 0.000 < 0.05) \). In other words, hypothesis \( H02 \) was rejected. It suggested that the impact of the administrative services on faculty members’ research outputs were different.

To find out which of the disciplines were statistically significantly different from one another, a Dunn-Bonferroni test between pairs of disciplines (e.g. between art and medical) were carried out as the post-hoc test. As a result, there were very strong evidences of differences between the following pairs of disciplines: Art \( (Md=0) \) vs. Engineer \( (Md=8) \) with an effect size \( r=0.467 \), Art \( (Md=0) \) vs. Medical \( (Md=10) \) with an effect size \( r=0.531 \) and Business/Management \( (Md =1) \) vs. Medical \( (Md =10) \) with an effect size \( r=0.407 \). There was no evidence of a difference
between the other pairs. Figure 2 describes the distribution of the publication counts of administrators in the four disciplines during the three defined periods. The decline in the number of publications was more obvious in the medical field than in other three fields, whereas the impact resulted in the smallest changes in the art field.

Figure 2. Comparison of the impact on administrators in different disciplines

Comparison between universities in different ranking levels
Hypothesis H03 was tested by a Kruskal-Wallis test. The independent variable was the ranking levels of the universities. We classified the universities into three ranking levels according to their rankings on the U.S. News & World Report. For example, universities that were ranked at top one to ten were considered as universities at the first ranking level. The Kruskal-Wallis H test revealed the statistically non-significant differences in the impact of the administrative roles on the research performance of academic administrators across the three ranking levels ($\chi^2(2, n = 116) = 2.352, p = 0.309 > 0.05$). The resultant effect size $r$ was 0.218, indicating a small to medium effect size using Cohen’s (1988) criteria (Pallant, 2013). Administrators from the higher-ranking universities (i.e. ranking 1 to 10) recorded a higher median score ($Md=11$) than administrators from universities in other two ranking levels, which both recorded median values of 4. As displayed in Figure 3, the total number of publications produced by universities at all three ranking levels substantially decreased during the in-position period.

Figure 3. Comparison of the impact on administrators working in universities at different ranking levels

Comparison between different administrative roles
Considering the administrative role as a factor, a Mann-Whitney U test was performed to examine the differences between department deans and university presidents. The results showed no significant difference in the impact of the administrative roles on the research performance of the department deans and the university presidents ($U=954.5$, $z=-0.561,$
$p=0.575$. The resultant effect size $r$ equalled to 0.052, which would be considered a very small effect size using Cohen’s (1988) criteria. It suggested the impact of administrative services on department deans and university presidents were both massive (as shown in Figure 4) but not significantly different. The median impact of the administrative roles on the university presidents’ research productions was 4.5 compared to 2 on the department deans.

![Figure 4. Comparison of the impact on administrators serving different roles](image)

**Conclusion**

Academic administrators are central to the development of higher education. As a serial of demanding, stressful, and time-intensive positions, previous qualitative studies suggested that management positions have various negative effects on one’s personal life and professional career (Werkema, 2009). The present study has used statistical analysis to examine the impact of administrative roles on the research performance of academic administrators in a quantitative way. It has demonstrated that the administrative services significantly influenced faculty members’ research outputs. As for different disciplines, administrators in the medical field are affected more heavily by their management roles. Such impact appears to be larger on administrators who serve in the higher-ranking universities. Both university presidents and department deans are substantially influenced by the administrative services, while the impact on the university presidents is slightly stronger than on the department deans.

The above results are limited by the population sampled and may not be applicable to all academic institute settings. This study currently relies on the publication counts to measure the research performance of faculty members. Future research including the measurement of academic presentations and research impact may corroborate these findings. In addition, a more detailed comparison will be presented as well, such as the research performance considering the order of authors.

**References**


Authorship, inventorship and division of labor in innovative research: an analysis of paper-patent pairs

Philippe Mongeon
philippe.mongeon@umontreal.ca
École de bibliothéconomie et des sciences de l’information, Université de Montréal, Montréal (Canada)
Observatoire des sciences et des technologies (OST), Centre interuniversitaire de recherche sur la science et la
technologie (CIRST), Montréal (Canada)

Abstract
Collaboration has long become the norm in science, and research is performed by increasingly large teams. Collaboration can take diverse forms, bring together researchers from different fields, status, and skills, and may lead to different patterns of tasks division between team members. This complicates the attribution of credit to individual authors, undermines research evaluation processes, and potentially reduces the efficacy of the reward system of science. This work in progress investigates division of labor and among team members using discoveries that were disclosed in both a patent and a paper, both document thus forming a paper-patent pair (PPP). Since the inventor status on a patent is attributed, in principle, for specific types of contribution (namely, those of a conceptual nature), comparing authors and inventors lists of PPPs can provide new insights on the division of labor and credit attribution practices in scientific research. This work analyzes the relationship between the author and inventor lists of 1,568 PPPs from five disciplines. We find that the individuals who are both authors and inventors are polarized towards the first and last author position. However, there appears to be disciplinary differences in the distribution of inventors in the various positions of the authors list.

Conference Topic
Collaboration and division of labor
Authorship
Inventorship
Paper-patent pairs

Introduction
Science is increasingly collaborative (Wuchty, Jones, & Uzzi, 2007), which leads to an increase in the number of authors of scientific papers. The longer author list, but the diverse types of collaboration (Maienschein, 1993), team compositions (Smith, 1971), and ways tasks are divided among team members (Larivière et al., 2016) makes the attribution of credit and responsibility to individuals an increasingly complex task (Rennie, Yank, & Emanuel, 1997). This is especially true for authors who are in the middle of the list (Bennett & Taylor, 2003; Shapiro, Wenger, & Shapiro, 1994). In this context, it is very difficult to deduct the type and extent of individual authors’ contribution based on their position in the byline. This is problematic for researchers whose career advancement largely depends on the credit they obtain for their work (Birnholtz, 2006; Cronin, 1996). It also undermines the efficiency of the reward system of science which requires that the best researchers be properly identified and rewarded (Merton, 1957). The importance of these issues has led to many studies that investigated authorship practices, or more specifically the relationship between individual contribution and authorship using different datasets and using both quantitative and qualitative methods.

Research in some disciplines may sometimes lead to technological innovations for which the researchers may choose to file a patent application. Patents are sometimes used as indicators of the usefulness or the inventiveness of research (Meyer, 2003), and because patenting is often highly valued in research evaluations, inventorship becomes, like authorship, a way for researchers to build a reputation and advance their career (Desrochers et al., in press). The
resulting paper-patent pair (PPP) can be a useful tool to analyze the type and substantiality of the tasks performed by the individual authors, and on credit attribution practices within research groups. The value of PPPs for investigating authorship practices lies in the fact that publishing and patenting operate in different realms and come with their distinct sets of rules and norms (Packer & Webster, 1996). For example, not all publishable research findings are patentable, and vice-versa. Similarly, contributions typically rewarded with authorship (e.g., technical contributions, data collection, drafting of the manuscript) are often not sufficient to be named inventor on a patent (Haeussler & Sauermann, 2013). Indeed, to be inventor, one must have conceived the invention (In re Hardee, 223 USPQ 1122, 1123; Comm’r Pat. 1984). Also, formulating desired outcomes is not sufficient: the inventor must have developed a non-obvious and concrete solution (Ex parte Smernoff, 215 USPQ 545, 547; Bd. App. 1982). In this context, comparing the inventor list and the author list of a patent and a paper that report the same discovery may indicate which of the authors have performed the conceptual work, and which have performed the technical work or other types of task that do not, in principle, lead to inventorship.

A few previous studies analyzed the difference between the inventors and authors list of PPPs. Looking at a set of 40 PPPs, Ducor (2000) found that the first and last authors were more likely than middle authors to be named inventors. He also found that last authors were more often listed as inventors than first authors. Using survey data from 2000 life scientists who self-reported having both published and patented a research finding, Haeussler and Sauermann (2013) found that contributions that are conceptual in nature are strong predictors of both authorship and inventorship, but that other types of contributions lead to authorship but not necessarily to inventorship. They also found that past scientific accomplishments is a predictor of authorship and that hierarchical position is a predictor of inventorship, but not authorship. Our study complements the previous research by identifying PPPs using worldwide paper and patent data covering all discipline for a large-scale comparison of authorship and inventorship practices. We analyze the relationship individual researchers’ position in the author list and their inclusion in the inventors list of the associated patent. More specifically, the present paper provides answers to the following research questions:

**RQ1.** What is the difference between the average number of authors and number of inventors?

**RQ2.** What is the relationship between the position in the authors’ list and the probability of also being listed as an inventor on the patent?

While these two questions are the focus of the preliminary results presented in this paper, they will be complemented by the following third research question in further steps of the study:

**RQ3.** What is the relationship between the authors’ characteristics, such as academic age, gender, productivity and scientific impact on the probability of also being listed as an inventor on the patent?

**Data and methods**

We collected all patent applications filed between 2001 and 2013 at the United States Patents and Trademark Office, and all articles published between 2000 and 2015 and indexed in Clarivate Analytics’ Web of Science (WoS). The articles published before 2000 are not considered since under the United States patent law, inventions published more than 12 months prior to the patent application are considered to be part of the public domain and are therefore
not patentable. Also, only articles for which the abstract is available were included, as the abstract is used for the pairing of the papers and patents.

To form PPPs, we first paired all articles and patents with at least one author-inventor match, based on the last name and first initial, which yielded more than 20 million potential pairs. Then, similarly to previous work by Magerman, Van Looy and Debackere (2015), we measured the content similarity between the articles (title and abstract) and patents (title and description). The difference between our method and that of Magerman et al. (2015) is that we used noun phrases instead of words to calculate the similarity between each PPPs. For both documents forming a potential pair, we divide the number of matching noun phrases by the total number of noun phrases in the title and abstract. The average of these two values is the similarity score of the pair.

To validate the matched pair, we manually verified all cases where the first name of the author was available in the WoS record. A match is considered a true positive when the first name of the author matches the first name of the inventor. The results of this validation are presented in Table 1.

Table 1. Validation of the author-inventor matches for cases where the full first names of the authors are available

<table>
<thead>
<tr>
<th>Similarity range</th>
<th>Number of author-inventors matches</th>
<th>True positives</th>
<th>False positives</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ .90</td>
<td>31</td>
<td>31</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>.80-.89</td>
<td>177</td>
<td>176</td>
<td>1</td>
<td>0.6%</td>
</tr>
<tr>
<td>.70-.79</td>
<td>176</td>
<td>174</td>
<td>2</td>
<td>1.1%</td>
</tr>
<tr>
<td>.60-.69</td>
<td>821</td>
<td>770</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>.50-.59</td>
<td>1,116</td>
<td>1,085</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>.40-.49</td>
<td>1,873</td>
<td>1,854</td>
<td>19</td>
<td>1.0%</td>
</tr>
<tr>
<td>.30-.39</td>
<td>3,647</td>
<td>2,758</td>
<td>889</td>
<td>24.4%</td>
</tr>
<tr>
<td>.20-.29</td>
<td>2,360</td>
<td>1,421</td>
<td>939</td>
<td>39.8%</td>
</tr>
</tbody>
</table>

We see that very few false positives occurred for PPPs with a similarity score of 0.40 or above (less than 1% overall), while 24.4% and 39.8% of author-inventor matches were false positives at a 0.30-0.39 and 0.20-0.30 similarity ranges, respectively. Therefore, we chose to limit the following analysis to all PPPs with a similarity of 0.40 or above, including those for which the full names of the authors were not available. We manually verified each PPP to ensure that no false negative remained in the author-inventor matching, and then excluded from the analysis all PPPs for which at least one inventor is not in the authors list to ensure that all contributors are included in the analysis. When multiple patent matched with a single paper, two criteria were used to select a single best matching patent: 1) we kept the pair with the highest number of authors who are also inventors on the patent, and 2) we keep the PPPs with the highest similarity score. We also limited the analysis to articles with 2 or more authors since the goal is to analyze division of labor and credit attribution among team members. One outlier, a paper with 161 authors, was also excluded from the analysis. The final dataset comprises 1,568 PPPs from five disciplines (see Table 2 in the result section).

Results

Table 2 presents the number of PPPs found for each discipline, the average number of authors, the average number of inventors, and the average inventors/authors ratio. Note that a small number PPPs were found in some of the fields not figuring in table 2 (e.g., Biology, Professional
Fields and Earth and Space). We excluded these PPPs from the analysis because their number was too small. Engineering and Technology, perhaps unsurprisingly, is the discipline with the most PPPs (N= 522) while the number of PPPs found for other disciplines ranges between 229 in Clinical Medicine to 279 in Biomedical Research. The results show that the average number of authors varies more between disciplines than the average number of inventors. Thus, the inventors/authors ratio ranges from 47.2% in Clinical Medicine, to 81.3% in Engineering and Technology, which suggests that important disciplinary differences exists in terms of authorship practices and division of labor among members of the research teams.

Table 2. Average number of authors, number of inventors, and inventors/authors ratio by discipline

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Number of PPPs</th>
<th>Avg. number of authors</th>
<th>Avg. number of inventors</th>
<th>Avg. inventors/authors ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biomedical Research</td>
<td>279</td>
<td>5.96</td>
<td>2.62</td>
<td>55.7%</td>
</tr>
<tr>
<td>Chemistry</td>
<td>267</td>
<td>4.53</td>
<td>2.39</td>
<td>62.8%</td>
</tr>
<tr>
<td>Clinical Medicine</td>
<td>229</td>
<td>6.06</td>
<td>2.24</td>
<td>47.2%</td>
</tr>
<tr>
<td>Engineering and Technology</td>
<td>522</td>
<td>3.35</td>
<td>2.56</td>
<td>81.3%</td>
</tr>
<tr>
<td>Physics</td>
<td>271</td>
<td>3.92</td>
<td>2.62</td>
<td>74.3%</td>
</tr>
<tr>
<td>Total</td>
<td>1,568</td>
<td>4.51</td>
<td>2.51</td>
<td>67.4%</td>
</tr>
</tbody>
</table>

We now investigate the relationship between an individual’s position in the author list and the likelihood also being listed as an inventor on the patent. If we divide the authors into three groups (first authors, middle authors, and last authors), we see that the inventors/authors ratio is smaller for middle authors than for first and last authors in all disciplines (Figure 1). We also observe that neither the first, or last authors are systematically listed as inventors on the matching patents. Moreover, last authors are most likely to be inventors than first authors in Biomedical Research, Clinical Medicine and Chemistry (although the difference seems negligible in the case of Chemistry), while in the first authors are more likely than last authors to be inventors in Physics and in Engineering and Technology.

![Figure 1. Inventors/authors ratio for first, middle and last authors](image)

As we have previously discussed in previous research (Mongeon, Smith, Joyal, & Larivière, 2016), the first author/middle author(s)/last author division of the byline does not adequately reflect the division of tasks among team members. The authors who are closer to the first and last position have typically contributed more substantially to the work than those listed further in the middle. Thus, Figure 2 displays the inventors/authors ratio as a function of the number...
of position separating authors from the first or last position. For articles with an odd number of authors, the middle author is duplicated and included in both the left and right side of the middle. For example, the third author of an article with five authors is included in both the “2 to first” and “2 to last” groups. The figure shows that the further an author is from the first and last positions, the least likely he or she is to appear as an inventor on the patent. We also see that in Chemistry, Engineering and Technology, and Physics the inventors/authors ratio tends to remain higher than in Biomedical Research and Clinical Medicine as we move further away from the first and last positions.

Note: the “middle” category includes all authors who are five positions or more (up to 14) away from the first or last position.

Figure 2. Inventors/authors ratio as a function of the number of position separating authors from the first or last position in the byline.

Discussion and conclusion

Our preliminary results display both similarities and differences between disciplines regarding the relationship between authorship and inventorship. Like Ducor (2000), we find that the first and last authors are more likely to appear in the inventors list in all discipline. This is in line with the previous finding that the first and last positions are the most important, and that middle authors have most likely made less substantial or technical contributions (Pontille, 2006).

In Biomedical Research and Clinical Medicine, the last authors are more often inventors than the first authors. Since the last author position is a senior scientist with a supervisory role and the first author is usually a junior scientist leading the experimental work (Pontille, 2006), this seems to confirm Haeussler and Sauermann’s (2013) finding that hierarchical position predicts inventorship. Interestingly, we find the opposite trend in Engineering and Technology and in Physics. This suggests either that 1) hierarchical position is a weaker predictor of inventorship in these fields, 2) that the author order in these field does not strictly follow the pattern described by Pontille (2006), or 3) that different patterns of division of labor are at play. This last hypothesis is also supported by the fact that middle authors are also more likely to be inventors in Engineering and Technology and Physics than in Biomedical Research and Clinical
Medicine. Chemistry appears to be somewhat in the middle, since the inventors/authors ratio is slightly higher for last authors compared to first authors. This ratio is also higher for middle authors in Chemistry than in Biomedical Research and Clinical Medicine, but lower than in Engineering and Technology and Physics. Further work will be needed to better understand these disciplinary differences.

Further developments of this research include the extension of the analysis to include data over a longer period. This will provide insights on the evolution over time of the authorship and inventorship practices across disciplines. The number of PPPs analyzed might also be further increased by manually identifying valid pairs for which the similarity score is below 0.40. Furthermore, the next step of the study will include contribution data when available, which will allow the comparison of the reported contribution with the information provided by the inventors list. Finally, researchers’ characteristics, such as the academic age, prominence (scientific output and impact), and gender will be included in the analysis to measure their potential effect on the patterns observed.

Acknowledgments

The author would like to thank Nees Jan van Eck from the Centre for Science and Technology Studies (CWTS) for the extraction of noun phrases from patents and publications. This research was funded by the Social Sciences and Humanities Research Council of Canada and by the Canada Research Chair on the Transformations of Scholarly Communication.

References


Martinez, & R. Wooley (Eds.), *STI2016 - Proceedings of the 21ST international conference on science and technology indicators* (pp. 448–452). Valencia, Spain, Spain.


Abstract
To position library and information science in the field of reading research, and to find the direction for endeavor, as well as support of theoretical and research methods, papers published between 2010 and 2016 in 6 journals on reading research in Journal Citation Reports (JCR) are collected and analyzed by means of CiteSpaceⅢ. The important authors, core journals, the main institutions and countries, research hotspots, research fronts and knowledge base, and pivotal documents in the field of reading research are identified. It is found that people of education and educational psychology have developed abundant theoretical and empirical studies, methods for tests and experiments, standards for assessment and instruction, as well as theories and tools developed. It is suggested that people of library and information science should forwardly seek opportunities to cooperate with reading experts and practitioners of education and educational psychology, and to expand their territory not only by applying theories and methods of reading research to practice, but also by actively engaging in study so as to offer better understanding about people’s reading behavior, enrich the content in this field, and to do more personalized reading guidance.

Conference Topic
Reading guidance; reading promotion; reading research; position of library and information science in reading research; CiteSpaceⅢ

Introduction
With the development of economy, science and technology, the overall national strength of China has greatly enhanced. Both the society and the government begin to attach importance to civil quality, and reading has been regarded the essential approach to improve citizen’s quality. Library Society of China released a declaration of library service in 2008 and claims that reading promotion is the basic goal of library service (Library Society of China (2008). However, although reading promotion has become mainstream library service in China, there is a lack of reading research (Fan, 2014).

On the other hand, reading environment has changed tremendously and people’s reading behavior has been changing accordingly. Increasing amount of time is spent on electronic documents and a screen-based reading behavior is emerging. More time spent on browsing and scanning, keyword spotting, one-time reading, non-linear reading, and people are reading more selectively, while less time is spend on in-depth reading and concentrated reading. Daily reading has become digital reading. (Fan 2014; Liu, 2005; Sharmin, Spakov & Raiha, 2015; Li, Guo, Chen & Luo, 2015; Kurata, Ishita, Miyata & Minami, 2017).

In this situation, it is necessary to figure out the big picture of reading research in a whole, so as to properly position library and information science (LIS hereafter) in this field and do
something to solve problems confronting both China and the field as well. However, reading is a research topic of long history, and there are an immense number of documents in this field. So Citesepe has been used to do data mining and visualization on reading research. For example, Yu and Mu (2011) analyzed reading research in China, Liu and Hua (2012) investigated children reading research, Feng and Qiao (2013) tried to find the hot keywords in reading research, Wei (2015) studied digital reading, Wang & Yuan (2016) inspected eye movements studies of reading behavior, Wu, Li and He (2017) made a comprehensive perspective on reading research in recent 10 years with data from SSCI/A&HCI/SCI-EXPANDED and CNKI core journals.

Usually there are specialized journals in any research field, and papers published in these journals always can reveal the development, changes, hotspots and fronts of the field. Up to now no such analysis has been done on journals of reading research, so this study will take journals in JCR (Journal Citation Reports) as data source and use CiteSpace as a tool, trying to dig deeper and find out more about this field as reference for LIS.

Data and methods

There are 6 journals on reading research in JCR in 2015, which are all indexed by citation database SSCI. Table 1 shows the basic information about these journals.

<table>
<thead>
<tr>
<th>Journal Title</th>
<th>Start Year</th>
<th>Country of Publication</th>
<th>Research Orientation</th>
<th>Impact Factor</th>
<th>Numbers of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scientific Studies of Reading</td>
<td>2003</td>
<td>England</td>
<td>Education &amp; Educational Research; Psychology</td>
<td>2.745</td>
<td>192</td>
</tr>
<tr>
<td>Reading Research Quarterly</td>
<td>1965</td>
<td>USA</td>
<td>Education &amp; Educational Research; Psychology</td>
<td>2.087</td>
<td>162</td>
</tr>
<tr>
<td>Reading and Writing</td>
<td>1990</td>
<td>Netherlands</td>
<td>Education &amp; Educational Research; Psychology</td>
<td>1.308</td>
<td>503</td>
</tr>
<tr>
<td>Journal of Research in Reading</td>
<td>2003</td>
<td>England</td>
<td>Education &amp; Educational Research; Psychology</td>
<td>0.917</td>
<td>190</td>
</tr>
<tr>
<td>Reading Teacher</td>
<td>1956</td>
<td>USA</td>
<td>Education &amp; Educational Research</td>
<td>0.697</td>
<td>618</td>
</tr>
<tr>
<td>Reading &amp; Writing Quarterly</td>
<td>2010</td>
<td>England</td>
<td>Education &amp; Educational Research</td>
<td>0.452</td>
<td>143</td>
</tr>
</tbody>
</table>

As the data of Reading & Writing Quarterly appears in SSCI as late as 2010, so set time range from 2010 to 2016, and take the titles of the 6 journal as search terms and search it in “title of publication” field in Web of Science, getting 192, 162, 503, 190, 618, 143 items from each journal respectively, 1808 items in total. The retrieval date is April 6, 2017. CiteSpace III 4.0.R5 was used to analyze these data. Due to limited space, co-citation network maps will not be included, and the results will be demonstrated in tables.

Results analysis

Institutions and countries

Choose “country” for “network node”, keep the other parameters unchanged, run CiteSpace III, and a national and regional co-citation network map comes out. Sorting the results by centrality, one can get the top 10 countries, as shown in Table 2. There are 6 countries with a
centrality bigger than 0.1 with USA as the leader. USA also has the largest publications. Though the publications of England, Germany, Australia, Spain and Netherlands are not very high, they have high quality. In CiteSpace, centrality means betweenness centrality.

Table 2. Top 10 countries in centrality

<table>
<thead>
<tr>
<th>No.</th>
<th>Centrality</th>
<th>Country</th>
<th>publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.77</td>
<td>USA</td>
<td>1126</td>
</tr>
<tr>
<td>2</td>
<td>0.37</td>
<td>ENGLAND</td>
<td>109</td>
</tr>
<tr>
<td>3</td>
<td>0.25</td>
<td>GERMANY</td>
<td>51</td>
</tr>
<tr>
<td>4</td>
<td>0.19</td>
<td>AUSTRALIA</td>
<td>48</td>
</tr>
<tr>
<td>5</td>
<td>0.18</td>
<td>SPAIN</td>
<td>27</td>
</tr>
<tr>
<td>6</td>
<td>0.15</td>
<td>NETHERLANDS</td>
<td>54</td>
</tr>
<tr>
<td>7</td>
<td>0.09</td>
<td>CANADA</td>
<td>113</td>
</tr>
<tr>
<td>8</td>
<td>0.08</td>
<td>NEW ZEALAND</td>
<td>19</td>
</tr>
<tr>
<td>9</td>
<td>0.05</td>
<td>FRANCE</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>0.03</td>
<td>ISRAEL</td>
<td>55</td>
</tr>
</tbody>
</table>

Select “institution” for “network node”, keep the other parameters unchanged, run CiteSpace III, one can get an institutions co-citation network map. Sorting the results by centrality, one can get the top 10 institutions, as shown in Table 3. There are only 4 institutions with a centrality no less than 0.1. Among the top 10 Institutions, 8 are from USA, and the other 2 are from Canada and Israel respectively.

Table 3. Top 10 institutions in centrality

<table>
<thead>
<tr>
<th>No.</th>
<th>Centrality</th>
<th>Institution</th>
<th>Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.22</td>
<td>Arizona State University</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>0.16</td>
<td>Michigan State University</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>0.15</td>
<td>Florida State University</td>
<td>71</td>
</tr>
<tr>
<td>4</td>
<td>0.14</td>
<td>University of Toronto</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>0.09</td>
<td>Vanderbilt University</td>
<td>36</td>
</tr>
<tr>
<td>6</td>
<td>0.07</td>
<td>Ohio State University</td>
<td>23</td>
</tr>
<tr>
<td>7</td>
<td>0.06</td>
<td>University of Haifa</td>
<td>28</td>
</tr>
<tr>
<td>8</td>
<td>0.06</td>
<td>University of Pittsburgh</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>0.06</td>
<td>University of North Carolina</td>
<td>20</td>
</tr>
<tr>
<td>10</td>
<td>0.05</td>
<td>Georgia State University</td>
<td>20</td>
</tr>
</tbody>
</table>

Influential authors

In CiteSpace III, select “Cited-Author” for “network node”, “title/abstractdescriptor /identifiers” for “Term Source”, choose “pathfinder” as the path search algorithm, set data extraction object as top 30, select the default threshold value, and choose “Burst Terms” for “Term Type”. Set “Time Scaling” for “2010-2016”, and set time partition as “1” year for each partition. Run CiteSpace III and get the author analysis map. Sorting the results by centrality, one can get the top 10 authors as listed in Table 4. There are 7 authors with a centrality bigger than 0.1. These authors are basically from fields of education and psychology. For example, Catherine McBride-Chang is a professor of developmental psychology at Chinese University.
of Hong Kong specializing in the acquisition of early literacy skills. Marilyn Jager Adams is a specialist in cognition and education, and she has developed a software using automated speech recognition to support learning to read. Charles Perfetti is the director of the Learning and Research Development Center at the University of Pittsburgh, and his research is centered on the cognitive science of language and reading processes, trying to develop a richer understanding of how language is processed in the brain.

### Table 4. Key authors in reading research

<table>
<thead>
<tr>
<th>No.</th>
<th>Centrality</th>
<th>Times Cited</th>
<th>Cited Author</th>
<th>Background/Interests</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.20</td>
<td>195</td>
<td>Wagner RK</td>
<td>psychology of reading</td>
</tr>
<tr>
<td>2</td>
<td>0.15</td>
<td>128</td>
<td>McBride-Chang C</td>
<td>developmental psychology</td>
</tr>
<tr>
<td>3</td>
<td>0.14</td>
<td>132</td>
<td>Adams MJ</td>
<td>cognition and education</td>
</tr>
<tr>
<td>4</td>
<td>0.14</td>
<td>93</td>
<td>Carlisle JF</td>
<td>dyslexia and applied psycholinguistics</td>
</tr>
<tr>
<td>5</td>
<td>0.13</td>
<td>171</td>
<td>Perfetti CA</td>
<td>cognitive science of language and reading processes</td>
</tr>
<tr>
<td>6</td>
<td>0.12</td>
<td>83</td>
<td>Biemiller A</td>
<td>human development and applied psychology/child study</td>
</tr>
<tr>
<td>7</td>
<td>0.10</td>
<td>83</td>
<td>Mccutchen D</td>
<td>phonology and morphology on students’ writing improvement</td>
</tr>
<tr>
<td>8</td>
<td>0.09</td>
<td>202</td>
<td>Stanovich KE</td>
<td>human development and applied psychology</td>
</tr>
<tr>
<td>9</td>
<td>0.09</td>
<td>132</td>
<td>Cain K</td>
<td>reading comprehension and children</td>
</tr>
<tr>
<td>10</td>
<td>0.09</td>
<td>115</td>
<td>Kintsch W</td>
<td>psychology and neuroscience</td>
</tr>
</tbody>
</table>

**Important journals**

Select “Cited Journal” for “network node”, keep the other parameters unchanged, run CiteSpace III, a journal co-citation network map appears. Based on the data, one can get the top 10 journals in terms of centrality, as shown in Table 5.

### Table 5. Important journals in reading research

<table>
<thead>
<tr>
<th>No.</th>
<th>Centrality</th>
<th>Times Cited</th>
<th>Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.26</td>
<td>944</td>
<td>Journal of Educational Psychology</td>
</tr>
<tr>
<td>2</td>
<td>0.19</td>
<td>240</td>
<td>Learning Disabilities</td>
</tr>
<tr>
<td>3</td>
<td>0.15</td>
<td>868</td>
<td>Reading Research Quarterly</td>
</tr>
<tr>
<td>4</td>
<td>0.12</td>
<td>752</td>
<td>Reading and Writing</td>
</tr>
<tr>
<td>5</td>
<td>0.08</td>
<td>223</td>
<td>Memory &amp; Cognition</td>
</tr>
<tr>
<td>6</td>
<td>0.07</td>
<td>663</td>
<td>Scientific Studies of Reading</td>
</tr>
<tr>
<td>7</td>
<td>0.07</td>
<td>393</td>
<td>Applied Psycholinguistics</td>
</tr>
<tr>
<td>8</td>
<td>0.07</td>
<td>270</td>
<td>American Educational Research Journal</td>
</tr>
<tr>
<td>9</td>
<td>0.07</td>
<td>174</td>
<td>Brain and Language</td>
</tr>
<tr>
<td>10</td>
<td>0.07</td>
<td>117</td>
<td>School Psychology Review</td>
</tr>
</tbody>
</table>

The journals are from psychology, cognitive science, education and learning science, brain science, as well as reading studies. There are only 4 journals with a centrality bigger than 0.1, however half are journals specialized on reading, i.e., *Reading Research Quarterly*, and *Reading and Writing*. 
Research hotspots

Select “Keyword” for “network node”, keep the other parameters unchanged, run CiteSpace III, a hotspot co-citation network map composed of keywords appears. List the results by centrality and citation frequency of keywords in descending order. Top 10 keywords are listed in Table 6 and there are 6 with a centrality bigger than 0.1. Children and comprehension take the top 2 positions. Dyslexia is always one of the hot topics of brain science and educational psychology.

It’s easy to find out from the table that hotspot keywords in this field are all basic keywords in the field of reading research. From Table 1 one can see that the directions of these 6 journals are education & educational research, and psychology. This means that most scholars publish paper in these journals usually have a background of education and psychology, so they study the basic problems in reading from its’ own perspective.

<table>
<thead>
<tr>
<th>No.</th>
<th>Centrality</th>
<th>Keyword</th>
<th>Times Cited</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.42</td>
<td>Children</td>
<td>487</td>
</tr>
<tr>
<td>2</td>
<td>0.21</td>
<td>Comprehension</td>
<td>280</td>
</tr>
<tr>
<td>3</td>
<td>0.21</td>
<td>Dyslexia</td>
<td>182</td>
</tr>
<tr>
<td>4</td>
<td>0.15</td>
<td>Strategy</td>
<td>161</td>
</tr>
<tr>
<td>5</td>
<td>0.12</td>
<td>Skill</td>
<td>208</td>
</tr>
<tr>
<td>6</td>
<td>0.12</td>
<td>Vocabulary</td>
<td>164</td>
</tr>
<tr>
<td>7</td>
<td>0.09</td>
<td>Method</td>
<td>116</td>
</tr>
<tr>
<td>8</td>
<td>0.08</td>
<td>Acquisition</td>
<td>190</td>
</tr>
<tr>
<td>9</td>
<td>0.08</td>
<td>Classroom</td>
<td>56</td>
</tr>
<tr>
<td>10</td>
<td>0.07</td>
<td>Motivation</td>
<td>75</td>
</tr>
</tbody>
</table>

Cluster Analysis on Keywords

After the hotspot co-citation network map is formed, select “find cluster” to do cluster analysis, and then choose “label clusters with title terms” to name each cluster. Set the threshold value of 50 and run cluster analysis on keywords. The value of M (modularity) is 0.4612, all the value of Silhouette are bigger than 0.5, and this means that the clustering result is quite good. Clustering results on keywords are listed in Table 7. In CiteSpace III, Silhouette is an indicator reflecting the homogeneity of members in the same cluster; the bigger the value, the higher the homogeneity, and the better the clustering results.

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Size</th>
<th>Silhouette</th>
<th>Year</th>
<th>Top Terms(log-likelihood)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#0 Common Core</td>
<td>32</td>
<td>0.893</td>
<td>2012</td>
<td>reality; classroom reading; romance</td>
</tr>
<tr>
<td>#1 Oral Reading Prosody</td>
<td>30</td>
<td>0.671</td>
<td>2010</td>
<td>SLI(specific language impairment); dyslexia; phonological awareness</td>
</tr>
<tr>
<td>#2 Bilingual Children</td>
<td>26</td>
<td>0.561</td>
<td>2010</td>
<td>reading comprehension; relationship; visual-word recognition</td>
</tr>
</tbody>
</table>

There are 3 clusters, i.e., common core, oral reading prosody and bilingual children. Common Core is a set of high-quality academic standards in mathematics and English language...
In cluster analysis, citing documents of high coverage reflect the research front, and cited documents with high centrality embody the knowledge base. In CiteSpace III, coverage means the percentage of documents in one cluster that the citing document cited. For example, if document A cites 15% of documents in cluster #1, then there will be a “0.15” in front of the title of document A, and the coverage of document A is 0.15. To understand the research front of this field, citing documents with a coverage bigger than 0.25 in cluster #0 are listed in Table 8. The document with the biggest coverage is about using smartphones to supplement classroom reading, reflecting the characteristic of mobile reading. Other documents are about guided reading, brain movies, independent reading, differentiated instruction, picturebooks, visual representations, common core, self-questioning, literature of modern family and reading in elementary school, of which some are new topic, and some are basic problems with new point of view of this time.

Table 8. Front documents with a coverage bigger than 0.25 in cluster #0

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Author(Year)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.44</td>
<td>Bromley, K (2012)</td>
<td>Using smartphones to supplement classroom reading</td>
</tr>
<tr>
<td>0.44</td>
<td>Fountas, IC (2012)</td>
<td>Guided reading: the romance and the reality</td>
</tr>
<tr>
<td>0.38</td>
<td>Wilson, D (2012)</td>
<td>Training the mind's eye &quot;brain movies&quot; support comprehension and recall</td>
</tr>
<tr>
<td>0.34</td>
<td>Sanden, S (2012)</td>
<td>Independent reading perspectives and practices of highly effective teachers</td>
</tr>
<tr>
<td>0.34</td>
<td>Watts-Taffe, S (2012)</td>
<td>Differentiated instruction: making informed teacher decisions</td>
</tr>
<tr>
<td>0.31</td>
<td>Martens, P (2012)</td>
<td>Learning from picturebooks: reading and writing multimodally in first grade</td>
</tr>
<tr>
<td>0.28</td>
<td>Coleman, JM (2012)</td>
<td>Visual representations in second graders' information book compositions</td>
</tr>
<tr>
<td>0.28</td>
<td>McLaughlin,, M (2012)</td>
<td>The common core: insights into the k-5 standards</td>
</tr>
<tr>
<td>0.25</td>
<td>Meese,, RL (2012)</td>
<td>Modern family: adoption and foster care in children's literature</td>
</tr>
<tr>
<td>0.25</td>
<td>Rockwell,, E (2012)</td>
<td>Appropriating written French: literacy practices in a Parisian elementary classroom</td>
</tr>
</tbody>
</table>

In the exported data of cluster analysis, only one keyword of each cited document is displayed. All the value of centrality, displayed keyword, and cluster ID of cited documents with centrality bigger than 0.1 are listed in Table 9.
Table 9. Cited documents in cluster analysis

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Centrality</th>
<th>Displayed Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>#0</td>
<td>0.19</td>
<td>strategy</td>
</tr>
<tr>
<td>#0</td>
<td>0.11</td>
<td>method</td>
</tr>
<tr>
<td>#1</td>
<td>0.13</td>
<td>acquisition</td>
</tr>
<tr>
<td>#1</td>
<td>0.11</td>
<td>skill</td>
</tr>
<tr>
<td>#1</td>
<td>0.11</td>
<td>developmental dyslexia</td>
</tr>
<tr>
<td>#2</td>
<td>0.34</td>
<td>comprehension</td>
</tr>
<tr>
<td>#2</td>
<td>0.19</td>
<td>vocabulary</td>
</tr>
<tr>
<td>#2</td>
<td>0.12</td>
<td>children</td>
</tr>
</tbody>
</table>

From Table 9 one can see the knowledge base of reading research, in which comprehension, vocabulary, strategy, acquisition, children, skill, method, developmental dyslexia are fundamental problems.

**Pivotal Documents**

In co-citation network map, the node documents with big centrality are often considered key documents that play the role of “pivotal point” in the development of domain knowledge (Chen, 2004, 2006). In CiteSpace III, select “cited reference” for “network node”, keep the other parameters unchanged, run CiteSpace III, the document co-citation network map will appear. Sort the results by descending order and list the documents with a centrality no less than 0.1, one can get the 10 pivotal documents, as showed in Table 10.

These pivotal documents are mostly published around 2005-2008, the newest appears in 2011. Perfetti is famous for his research in reading ability, and this literature review is based on his Distinguished Research Award addressed to the Society for the Scientific Study of Reading in 2006. According to lexical quality hypothesis (LQH), variation in the quality of word representations has consequences for reading skill, including comprehension. High lexical quality allows for rapid and reliable meaning retrieval, and low-quality representations lead to specific word-related problems in comprehension. In the first paper, Perfetti (2007) reviews 6 lines of research on adult readers which demonstrate some of the implications of the LQH and provide evidence that word-level knowledge has consequences for word meaning processes in comprehension.

The Simple View of Reading claims that reading comprehension is the product of two processes: word recognition and linguistic comprehension. Word recognition refers to the ability to read printed words without the aid of context, and linguistic comprehension refers to the ability to understand language. In the 2nd paper, Adlof, Catts, & Little (2006) reported their experimental study which examined the independent contribution of fluency to reading comprehension, and investigated whether a separate fluency component should be added to the Simple View of Reading. Reading and language measures were administered to 604 children in 2nd, 4th, and 8th grades. Results from both large group structural equation modeling analyses and individual profile analyses indicate that fluency does not independently contribute to reading comprehension, so the Simple View of Reading does not need to be modified to include a separate fluency component.

In the 3rd paper Share (2008) systematically critique current reading research and practice which have confined reading science to Anglocentric research agenda addressing theoretical and applied issues with limited relevance for a universal science of reading.
<table>
<thead>
<tr>
<th>No.</th>
<th>Centrality</th>
<th>Times Cited</th>
<th>Author/Year</th>
<th>Title</th>
<th>Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.27</td>
<td>46</td>
<td>PERFETTI C/2007</td>
<td>Reading ability: Lexical quality to comprehension</td>
<td>Scientific Studies of Reading</td>
</tr>
<tr>
<td>2</td>
<td>0.21</td>
<td>18</td>
<td>ADLOF SM/2006</td>
<td>Should the Simple View of Reading Include a Fluency Component?</td>
<td>Reading and Writing</td>
</tr>
<tr>
<td>3</td>
<td>0.19</td>
<td>59</td>
<td>SHARE DL/2008</td>
<td>On the anglocentricities of current reading research and practice: The perils of overreliance on an &quot;Outlier&quot; orthography</td>
<td>Psychological Bulletin</td>
</tr>
<tr>
<td>4</td>
<td>0.19</td>
<td>60</td>
<td>ZIEGLER JC/2005</td>
<td>Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory</td>
<td>Psychological Bulletin</td>
</tr>
<tr>
<td>5</td>
<td>0.17</td>
<td>16</td>
<td>GRAHAM S/2011</td>
<td>Writing to Read: A Meta-Analysis of the Impact of Writing and Writing Instruction on Reading</td>
<td>Harvard Educational Review</td>
</tr>
<tr>
<td>6</td>
<td>0.13</td>
<td>22</td>
<td>KIRBY JR/2008</td>
<td>Can the simple view deal with the complexities of reading?</td>
<td>Literacy</td>
</tr>
<tr>
<td>7</td>
<td>0.12</td>
<td>21</td>
<td>KUHN MR/2010</td>
<td>Aligning Theory and Assessment of Reading Fluency: Automaticity, Prosody, and Definitions of Fluency</td>
<td>Reading Research Quarterly</td>
</tr>
<tr>
<td>8</td>
<td>0.12</td>
<td>35</td>
<td>NAGY W/2006</td>
<td>Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle-school students</td>
<td>Journal of Educational Psychology</td>
</tr>
<tr>
<td>9</td>
<td>0.10</td>
<td>8</td>
<td>BARCA L/2006</td>
<td>Italian developmental dyslexic and proficient readers: where are the differences?</td>
<td>Brain and Language</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>46</td>
<td>KEENAN JM/2008</td>
<td>Reading Comprehension Tests Vary in the Skills They Assess: Differential Dependence on Decoding and Oral Comprehension</td>
<td>Scientific Studies of Reading</td>
</tr>
</tbody>
</table>
In the 4th paper, based on review of the rich cross-language database concerning phonological development, reading development, and dyslexia, Ziegler & Goswami (2005) proposed a psycholinguistic grain size theory of reading and its development, trying to integrate these cross-language data into a theoretical framework for describing reading acquisition, skilled reading, and dyslexia in different language.

The 5th paper examines the effectiveness of writing as a tool for improving students’ reading. Graham & Hebert (2011) collected 752 documents on this topic of which 95 experiments met their inclusion criteria. By meta-analysis of these true and quasi-experiments, they present evidence that writing about material read improves students’ comprehension of it; that teaching students how to write improves their reading comprehension, reading fluency, and word reading; and that increasing how much students write enhances their reading comprehension. Their findings provide empirical support for long-standing beliefs about the power of writing to facilitate reading.

In the 6th paper Kirby & Savage (2008) review the Simple View of Reading model and examine its nature, applicability and validity. After systematic review on related studies, they conclude that the model serves as an abstract framework for understand the relationship between global linguistic comprehension and word-reading abilities in reading comprehension, however it is neither a full theory of reading nor a blueprint for instruction. They indicate several areas in which the model is incomplete or in need of further empirical support, i.e., the way word decoding is conceptualized, the way reading comprehension is measured, reading comprehension strategies, the role of reading fluency, reading with illustrations and second-language reading.

Aiming at the relatively narrow definition of reading fluency, which emphasized automatic word recognition, in the 7th paper, after reviewing multiple ways of conceptualizing reading fluency, Kuhn, Schwanenflugel, and Meisinger (2010) propose a definition to expand this understanding by synthesizing several key aspects of research on reading fluency, especially automaticity and prosody, so as to develop assessments and instruction to facilitate the readers’ construction of meaning.

The 8th paper reports the work of Nagy, Berninger, & Abbot (2006). They conducted an experimental study on morphology’s contribution to literacy outcomes of upper elementary and middle-school students, with 607 students in Grades 4 through 9 in a suburban school district as participants. They use structural equation modeling to evaluate the contribution of morphological awareness, phonological memory, and phonological decoding to reading comprehension, reading vocabulary, spelling, and accuracy and rate of decoding morphologically complex words for these students. The results are quite positive, with morphological awareness making a significant unique contribution to reading comprehension, reading vocabulary, and spelling for all students, to all measures of decoding rate for the 8th/9th-grade students, and to some measures of decoding accuracy for the 4th/5th-grade and 8th/9th-grade students. Morphological awareness also made a significant contribution to reading comprehension above and beyond that of reading vocabulary for all students.

In the 9th paper Barca, Burani, Di Filippo, and Zoccolotti (2006) reported an experimental study in which they tried to understand whether Italian developmental dyslexics make use of the whole-word reading procedure. They recruited 82 Italian 6th graders as the dyslexic group and 68 participants with normal reading ability as the control group. They investigated both lexical and nonlexical reading procedures in dyslexic children through word frequency effect and the effect of contextual grapheme-to-phoneme conversion rules. Although dyslexics were slower and less accurate than controls, they were affected by word frequency, grapheme contextuality, and their interaction in a similar manner as average readers. These results show the use of lexical reading in Italian dyslexics, and reject the claim of a deficit in whole-word
processing with consequent over-reliance on the non-lexical route. This is the first demonstration of the availability and use of lexical reading in Italian dyslexics.

Reading comprehension is a complex cognitive construct, consisting of multiple component skills. However, comprehension tests are focusing mainly on word decoding skills, and they are often used interchangeably, suggesting an implicit assumption that they are all measuring the same thing. In the 10th paper Keenan, Betjemann, and Olson (2008) compare some of the most popular reading comprehension measures used in research and clinical practice in the United States. Participants are 510 children from 27 different school districts in Colorado. Measures include reading comprehension, listening comprehension, word decoding and non-word decoding. The modest inter-correlations among the tests suggested that they were measuring different skills.

These documents are basically on the fundamental problems of reading research, especially theoretical issues. There are 2 papers on the Simple View of Reading, 2 papers on assessment, others are on typical topics such as developmental dyslexia, strategy of reading instruction, etc. People in this field have been trying to understand completely on the process of reading, to build a complete model to describe the process, and to design multiple strategies to promote reading comprehension.

**Discussion**

Reading research is a field of long history, and people of education and educational psychology are the main force. They have developed abundant theoretical and empirical studies, methods for tests and experiments, standards for assessment and instruction, as well as theories and tools.

People of LIS are mainly working in reading guidance and promotion, focusing on practice other than theoretical studies. This is why there is a lack of reading research.

This is a time of new media, and there are new opportunities for people in LIS on both reading guidance and reading research. On the one hand, they should absorb the mature theories, standards, methods and tools developed by people of education and educational psychology to do more elaborate reading promotion. On the other hand, they can actively cooperate with experts and practitioners of education and educational psychology to do reading research. In this way, they can play more important role in reading research as well practice. In fact the new reading phenomenon and reading behavior in the new information environment have become the hot topic of reading research, and the visibility of reading research doing by LIS is increasing day by day (Wu, Li & He, 2017).

Reading guidance and promotion of LIS mostly aim at reading happening in people’s spare time other than school time or work time. There is no doubt that LIS is playing an important role in enriching people’s reading life and enhancing their reading ability. Based on daily practice it is also convienient for LIS profession to study people’s reading behavior so as to get deeper understanding about reading and to provide service at more professional level. By engaging more in reading research in collaboration with other fields in the practice of reading guidance and prompting, LIS profession can not only apply theories and methods of this field to practice, but also offer better understanding about people’s reading behavior and enrich the content in reading research.

For example, repeated reading method is one of the most well-known methods for improving reading fluency. It is often considered a successful fluency intervention in improving both reading fluency and reading comprehension because it gives children practice recognizing the same words in their meaningful context. (Adlof, Catts, & Little 2006). However, the time students can devote to repeated reading is limited, and the effort everyone should make on repeated reading is different. So LIS profession has opportunities to develop and launch elaborate repeated reading projects in cooperation with teachers in schools, and with experts
on reading. In these projects data could be collected and analyzed to test the effectiveness of repeated reading method, and the results may be used to design more personalized reading plan for each participants.

Conclusions

A co-citation visualization analysis has been made with CiteSpace III on the documents of 6 reading research journals in JCR, which are all in citation database SSCI, between 2010-2016. It identifies the important authors, core journals, the main institutions and countries, research hotspots, research fronts and knowledge base, and pivotal documents in the field of reading research.

It is found that people of education and educational psychology are the main force in reading research. The abundant theoretical and empirical studies, methods for tests and experiments, standards for assessment and instruction, as well as theories and tools developed by them are all available for people of LIS to do more personalized reading guidance and deeper reading research.

It is suggested that people of LIS should forwardly seek opportunities to cooperate with reading experts and practitioners of education and educational psychology, and to expand their territory in this new media time not only by applying theories and methods of reading research to practice, but also by actively engaging in study so as to offer better understanding about people’s reading behavior and enrich the content in this field.

Acknowledgments

This study is funded by the National Social Science Foundation under grant No.13BTQ024.

References

Kirby, J. R., & Savage, R. S. (2008). Can the simple view deal with the complexities of reading? Literacy, 42(2), 75-82.


Feng Xiangmei & Qiao Huan. (2013). Analysis on Mapping Knowledge Domain on Hot Keywords of Reading Researches at Home and Abroad. *Information Research, 3*, 11-16.


Path to Success: An analysis of US educated elite academics in the United States

Tolga Yuret

tyuret@gmail.com
Istanbul Technical University, Istanbul (Turkey)

Abstract

Few academics achieve to become professors in prestigious universities. These elite academics need a very strong educational background in addition to high academic ability for this achievement. We have two main objectives in analyzing the educational backgrounds of the professors. First, educational institutions are evaluated in terms of the number of the alumni who become professors. Second, professors’ path to success is analyzed and some light is shed on issues such as academic mobility. The educational backgrounds of 14,310 full professors from top 48 universities in the United States are analyzed. This study is confined to undergraduate and graduate degrees that are attained from United States. The source of educational attainment among professors is found to be concentrated. 72 percent (44 percent) of the professors get their PhD (undergraduate) degrees from 20 universities. The academic mobility is found to be very high. Only 16 percent of the professors work in the same state as they get their undergraduate degrees.

Keywords: Academic rankings; academic mobility; inbreeding; financing of education; time to graduate

Conference Topic

Country-level studies, studies on the individual level scientists

Introduction

Thousands of candidates compete for few tenure-track positions in the prestigious universities each year. Many tenure-track faculty members cannot complete the tenure process with success. Some tenured associate professors fail to become tenured full professors. At the end, few PhD holders achieve to become full professors in the prestigious universities. Therefore, it is interesting to analyze the path to this considerable success.

There are many properties of the successful academics. They have good social skills which enable them to direct laboratories and write joint papers. They expend considerable time and effort to do their scientific analysis. They have high academic ability to solve difficult scientific problems. But all these properties would not be sufficient if the academics do not get a strong academic training. Therefore, educational backgrounds of the professors have the utmost importance in the path to success.

We have two main aims in analyzing the educational backgrounds of the professors. First, educational institutions are evaluated in terms of the number of alumni who work as professors in prestigious universities. Both undergraduate and graduate degrees are considered to provide two separate ranks for the institutions. Second, the path to become a professor is analyzed. In this regard, questions such as whether public or private institutions educate more professors for each other are answered. Moreover, mobility of the professors and the degree of inbreeding are measured. Lastly, the graduation time of the professors are computed.
The educational backgrounds of 14,310 full professors who work in top 48 universities in the United States are analyzed for this study. Professors who are from 16 academic fields in natural sciences, engineering, social sciences and humanities are considered. The analysis of this study is confined to the degrees attained from United States.

The preceding paper Yuret (2017) uses the same data-set as this paper but focuses on the professors who have foreign degrees. 34.5 percent of the undergraduate degrees and 12.5 percent of the PhD degrees are found to be from foreign institutions. The source of education is found to be largely concentrated in developed countries. The ratio of foreign-educated professors are found not to vary with the ranking of the university or the public ownership of the university.

In this paper, we find that professors attain their degrees from a relatively few number of universities. 72 percent (44 percent) of the professors get their PhD (undergraduate) degrees from 20 universities. Moreover, top 17 universities in terms of the PhD alumni who become professors are also among top 20 universities in terms of the undergraduate alumni who become professors.

The concentration of source of education among elite academics has been noted by previous studies. Stock et al. (2000) analyze the economics PhD alumni and notes that only 2 out of 178 alumni work in a university that is ranked higher than the university that they graduate. Majority of the PhD graduates work in universities ranked 50 below. Therefore, it is natural that the prestigious universities have professors graduated mostly from top institutions.

Clauset et al. (2015) analyze 19000 professors from business, computer science and history. They show that 71 to 86 percent of the professors attain their PhD degrees from the quarter of the institutions. This study is different from Clauset et al. (2015) in two respects. First, this paper has fewer universities but spans more academic fields. Second, this paper computes the concentration at both undergraduate and PhD levels.

The success of the graduates have been used as a measure for the academic evaluation of the universities. For example, one of the measures that Shanghai Rankings use is the number of alumni who get Nobel Prize or Fields Medal. Laband (1985) and Amir and Knauff (2008) rank economics departments in terms of the placement of their graduates.

We compare the states in which the professors work to the states that professors get their education. There are two objectives in doing this. First, academic mobility is compared to the general labor mobility. Only 16 percent of the professors are found to work in the state that they get their undergraduate degrees. In contrast, Kodrzycki (2001) finds that 60 percent of the people who have some graduate degree works in the state that they get their undergraduate degrees. Second, the degree that some states such as California educate professors for other states such as Texas is measured. 21 percent of the professors who work in California are found to get their PhD degrees from Texas whereas only 2 percent of the professors who work in Texas are found to get their PhD degrees from California.

It has been noted that inbreeding is a problem for some countries. For example, Inanc and Tuncer (2011) find that the inbred faculty have an H index 89 percent lower than the noninbred faculty in some prestigious universities in Turkey. However, it has been noted that inbreeding is not a problem for the United States. For example, Wyer and Conrad (1984) find that inbred faculty members are as productive as other faculty members in many respects by using a survey that covers 160 universities in the United States.

The degree of inbreeding is also measured in this study. The inbreeding is found to be very small. Only 4 percent of the professors work in the university that they get their undergraduate degrees. The ratio for the PhD degrees is also found to be 4 percent. The inbreeding ratio is below 10 percent for all of the universities except for Harvard and MIT.

There is a considerable pressure for the PhD students to graduate faster because a late graduate is considered as a negative signal to the prospective employers. We find that the median PhD
graduate who succeeds to become a professor graduates much faster than the median PhD graduate.

**Data**

The educational backgrounds of the elite academics are considered both in this study and in Yuret (2017). This study analyzes the degrees attained from institutions in the United States whereas Yuret (2017) analyzes the foreign degrees. The data for both studies are collected at the same time. Naturally, there are similarities in data descriptions of both studies.

48 universities from United States that are top 100 in 2015 Shanghai rankings are considered.¹

There are 51 universities from United States that are top 100. Three of these universities are excluded because they specialize on life and/or medical sciences.²

16 academic fields are chosen by using the following three criteria.

- a-) Selected fields exist in most universities.
- b-) Selected fields are mostly organized as departments.
- c-) It is possible to attain educational backgrounds of elite academics in selected fields.

At the end, three natural sciences (chemistry, mathematics, physics), seven engineering (bio/biomedical engineering, chemical engineering, civil and environmental engineering, electrical and computer engineering & computer science, industrial engineering, materials science and engineering, mechanical & aero engineering), four social sciences (economics, political science, psychology, sociology) and two humanities (history, philosophy) fields are selected. We have some composite fields such as electrical and computer engineering & computer science because we are unable to separate them for some universities.

The undergraduate and PhD institutions from which the professors graduate and their graduation dates are collected between September 2015 and January 2016. We rely mainly on internet sources. We use the official websites of the professors, internet search engines (e.g google), undergraduate catalogs, commencement leaflets, PhD data-bases, academic trees (e.g. mathematical genealogy), internet encyclopedias (e.g. wikipedia), field specific data-bases (e.g. Inspire), social networking sites (e.g. Linkedin) and other various sites. We also use reference books (e.g. American Men & Women in Science). We send e-mails to professors if we are unable to find their educational background information from these sources. Around 1/3 of the professors replied.

After all this search, complete educational background information about 95.9 percent of the 14,310 full professors are collected. 97.2 percent of the professors have only missing graduation date information. Almost all the missing data is about undergraduate degrees. Complete information about the PhD degrees for 99.7 percent of the professors is available.

The zip codes of the universities are collected from their official web pages and wikipedia. The coordinates for the zip codes are available from <federalgovernmentzipcodes.us>. The approximate distances between the zip codes are computed by using these coordinates.

**Ranking of the educational institutions by the number of alumni who become professors**

Table 1 lays out the concentration of the education of the professors. First, universities are sorted by the number of the alumni who works as professors within each field. Then, the ratio of professors who get their education in top education providers is computed. For example, 16 percent of all chemistry professors get their undergraduate degrees from top five universities that educate most professors in chemistry.

Table 1 shows that undergraduate education is much more dispersed than graduate education. When all fields are considered, 72 percent of the professors get their PhD degrees from 20 universities whereas the rate is only 44 percent for undergraduate degrees. 35 (20) percent of the professors get their PhD (undergraduate) degrees from 5 universities and 51 (31) percent of the professors get their PhD (undergraduate) degrees from 10 universities.
The concentration of all fields is lower than the concentration within fields. For example, 51 percent of the PhD degrees are from 10 universities when all fields are considered. However, more than 51 percent of the professors are from 10 universities in all fields except for bio-engineering and psychology. This is because the universities are specialized in different fields. For example, if University X educates all the professors in Field A and University Y educates all the professors in Field B, then each university educates 50 percent of all of the professors but educates 100 percent of the professors in their respective fields.

There is a considerable concentration of PhD degrees in almost all fields. More than 40 percent of the PhD degrees are from 5 universities, more than half of the PhD degrees are from 10 universities and more than 70 percent of the degrees are from 20 universities for all fields except for psychology and bio-engineering.

There is a variation of concentration among fields. For example, 88 percent of the economics professors are from 20 universities where the ratio is 59 percent for psychology professors. Psychology degrees are less concentrated than economics degrees at the undergraduate level as well. Chemistry is the least concentrated science field and bio-engineering is the least concentrated engineering field.

In Table 2, universities are ranked in terms of the number of undergraduate alumni who work as professors. Top 20 universities are listed. The rankings are sensitive to the choices that we have made. For example, we do not control for the size of the universities. Caltech may rank higher if a size control is applied. Moreover our selection of fields depend on factors other than the importance of the field. For example, only two humanities fields are included because of the data availability. If more humanities fields could be added, Yale might have ranked higher. It is interesting to note that the academic future is mainly decided at the undergraduate level. Although Harvard tops the list in three fields, there are few Harvard graduates who become engineering professors. Likewise, a student who is an MIT graduate is very unlikely to become a humanities or social sciences professor. There are approximately eight times more engineering professors who are graduated from MIT than UCLA although more UCLA graduates become social science professors than MIT graduates.

If we added one more university to the list in Table 2, then that university would be Swarthmore which is a liberal arts college. There are 83 professors who get their undergraduate education from Swarthmore. That is, one less professor gets undergraduate education from Swarthmore compared to Michigan State. This is a considerable success considering that we do not control for the size of the universities.
Table 1. Concentration of Undergraduate and PhD degrees

<table>
<thead>
<tr>
<th>All Fields</th>
<th>Undergraduate</th>
<th>PhD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top 5</td>
<td>Top 10</td>
</tr>
<tr>
<td>Chemistry</td>
<td>0.16</td>
<td>0.25</td>
</tr>
<tr>
<td>Mathematics</td>
<td>0.29</td>
<td>0.42</td>
</tr>
<tr>
<td>Physics</td>
<td>0.35</td>
<td>0.48</td>
</tr>
<tr>
<td>Natural Sciences</td>
<td>0.26</td>
<td>0.37</td>
</tr>
</tbody>
</table>

| Bio Eng    | 0.18  | 0.30   | 0.44   | 0.28  | 0.43   | 0.64   |
| Chem Eng   | 0.20  | 0.33   | 0.50   | 0.41  | 0.60   | 0.77   |
| Civ & Env  | 0.19  | 0.29   | 0.44   | 0.41  | 0.59   | 0.76   |
| ECE & CS   | 0.29  | 0.43   | 0.58   | 0.41  | 0.56   | 0.76   |
| Ind Eng    | 0.29  | 0.41   | 0.58   | 0.45  | 0.65   | 0.82   |
| Mat Sci Eng| 0.26  | 0.39   | 0.57   | 0.43  | 0.60   | 0.78   |
| Mech Aero  | 0.23  | 0.35   | 0.52   | 0.44  | 0.60   | 0.78   |
| Engineering| 0.23  | 0.34   | 0.49   | 0.39  | 0.53   | 0.71   |

| Economics  | 0.27  | 0.39   | 0.52   | 0.47  | 0.71   | 0.88   |
| Political Sci. | 0.19 | 0.30   | 0.43   | 0.40  | 0.62   | 0.82   |
| Psychology | 0.14  | 0.23   | 0.35   | 0.25  | 0.42   | 0.59   |
| Sociology  | 0.18  | 0.29   | 0.43   | 0.44  | 0.60   | 0.78   |
| Social Sciences | 0.17 | 0.27   | 0.41   | 0.33  | 0.53   | 0.72   |

Table 2. Undergraduate degrees of professors by field

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Natural Sciences</th>
<th>Engineering</th>
<th>Social Sciences</th>
<th>Humanities</th>
<th>Top 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harvard</td>
<td>536</td>
<td>195</td>
<td>78</td>
<td>170</td>
<td>93</td>
<td>184</td>
</tr>
<tr>
<td>MIT</td>
<td>466</td>
<td>147</td>
<td>275</td>
<td>37</td>
<td>7</td>
<td>154</td>
</tr>
<tr>
<td>Berkeley</td>
<td>336</td>
<td>94</td>
<td>135</td>
<td>82</td>
<td>25</td>
<td>73</td>
</tr>
<tr>
<td>Cornell</td>
<td>266</td>
<td>73</td>
<td>102</td>
<td>66</td>
<td>25</td>
<td>49</td>
</tr>
<tr>
<td>Princeton</td>
<td>250</td>
<td>78</td>
<td>74</td>
<td>53</td>
<td>45</td>
<td>80</td>
</tr>
<tr>
<td>Yale</td>
<td>226</td>
<td>54</td>
<td>55</td>
<td>62</td>
<td>55</td>
<td>57</td>
</tr>
<tr>
<td>Michigan</td>
<td>215</td>
<td>37</td>
<td>85</td>
<td>69</td>
<td>24</td>
<td>30</td>
</tr>
</tbody>
</table>
Table 3 ranks the universities in terms of the PhD alumni who become professors. Top 20 universities are listed. There is a close parallel between undergraduate and graduate rankings. Top 17 universities in Table 3 are also among top 20 universities in Table 2. However, there are also differences between these two listings. For example, Stanford ranks 3rd in Table 3 but ranks 8th in Table 2. On the contrary, Cornell ranks 4th in Table 2 but ranks 10th in Table 3. Table 3 shows that there is a high degree of specialization in the universities. Berkeley tops the list although they are the leader only in natural sciences. The leader of the engineering field is MIT and the leader in both humanities and social sciences fields are Harvard. The last column in Table 3 considers the professors from eight US universities that are among top ten in the world. Table 3 shows that the rankings are sensitive to the quality of the universities. Columbia which ranks 14th educates more professors for top ten universities than Michigan which ranks 6th.

<table>
<thead>
<tr>
<th>University</th>
<th>Total</th>
<th>Natural Sciences</th>
<th>Engineering</th>
<th>Social Sciences</th>
<th>Humanities</th>
<th>Top 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berkeley</td>
<td>1116</td>
<td>356</td>
<td>445</td>
<td>216</td>
<td>99</td>
<td>284</td>
</tr>
<tr>
<td>MIT</td>
<td>975</td>
<td>240</td>
<td>570</td>
<td>144</td>
<td>21</td>
<td>327</td>
</tr>
<tr>
<td>Stanford</td>
<td>847</td>
<td>178</td>
<td>420</td>
<td>188</td>
<td>61</td>
<td>239</td>
</tr>
<tr>
<td>Harvard</td>
<td>839</td>
<td>278</td>
<td>117</td>
<td>291</td>
<td>153</td>
<td>306</td>
</tr>
<tr>
<td>Princeton</td>
<td>603</td>
<td>253</td>
<td>121</td>
<td>120</td>
<td>109</td>
<td>176</td>
</tr>
<tr>
<td>Michigan</td>
<td>428</td>
<td>48</td>
<td>162</td>
<td>154</td>
<td>64</td>
<td>37</td>
</tr>
<tr>
<td>Chicago</td>
<td>402</td>
<td>151</td>
<td>22</td>
<td>163</td>
<td>66</td>
<td>74</td>
</tr>
<tr>
<td>Yale</td>
<td>399</td>
<td>89</td>
<td>52</td>
<td>140</td>
<td>118</td>
<td>71</td>
</tr>
<tr>
<td>Caltech</td>
<td>378</td>
<td>138</td>
<td>218</td>
<td>22</td>
<td>0</td>
<td>85</td>
</tr>
<tr>
<td>Cornell</td>
<td>373</td>
<td>136</td>
<td>155</td>
<td>50</td>
<td>32</td>
<td>54</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>370</td>
<td>85</td>
<td>115</td>
<td>116</td>
<td>54</td>
<td>33</td>
</tr>
<tr>
<td>Urbana</td>
<td>369</td>
<td>76</td>
<td>225</td>
<td>63</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>UCLA</td>
<td>326</td>
<td>78</td>
<td>111</td>
<td>83</td>
<td>54</td>
<td>40</td>
</tr>
<tr>
<td>Columbia</td>
<td>324</td>
<td>92</td>
<td>65</td>
<td>84</td>
<td>83</td>
<td>76</td>
</tr>
<tr>
<td>Minnesota</td>
<td>260</td>
<td>50</td>
<td>93</td>
<td>98</td>
<td>19</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 3. PhD degrees of the elite academics
The rankings provided in Table 2 and 3 evaluate the success of the alumni who have graduated many years before. These rankings serve two purposes. First, the rankings of the universities do not change much through time (Daly et al. 2006). Therefore, the success of past alumni is highly correlated with the current quality of the universities. Second, reputation of the alumni of the institutions is important. An alumni of a university that has educated many professors in prestigious universities may benefit from this reputation.

Path to Success: Public ownership and rankings of the universities
Public universities provide most of the human capital necessary for the universities in many countries. United States is one of the few exceptions. Table 4 shows that 70 percent of the professors in private universities are educated in private universities whereas only half of the professors in public universities are educated in public universities. The ratios are similar for both undergraduate and graduate degrees. The fact that private institutions educate professors for public institutions may be used as an argument for increasing subsidies for private universities.

<table>
<thead>
<tr>
<th>Undergraduate Professor in Private</th>
<th>PhD Professor in Private</th>
<th>Undergraduate Professor in Public</th>
<th>PhD Professor in Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private Education from Public</td>
<td>0.70</td>
<td>0.53</td>
<td>0.30</td>
</tr>
<tr>
<td>Total</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 5 shows that the relation between the Shanghai rank of the university that the professors get their education and Shanghai rank of the university that they currently work. It is clearly seen that the top institutions educate professors for the lower ranked institutions. For example, almost half of the professors who work in the universities that rank 11 to 25 get their PhD education from top 10 universities whereas only 14 percent of the professors who work in top 10 universities get their PhD from universities that rank 11 to 25. The dominance of the top 10 universities are weaker for the undergraduate degrees. Only a quarter of the professors who work in universities ranked 11 to 25 get their undergraduate degrees from top ten universities.

<table>
<thead>
<tr>
<th>Professors in</th>
<th>Top 10</th>
<th>11-25</th>
<th>26-50</th>
<th>51-75</th>
<th>76-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 10</td>
<td>0.41</td>
<td>0.26</td>
<td>0.23</td>
<td>0.19</td>
<td>0.13</td>
</tr>
<tr>
<td>11-25</td>
<td>0.14</td>
<td>0.18</td>
<td>0.15</td>
<td>0.14</td>
<td>0.10</td>
</tr>
<tr>
<td>Undergraduate</td>
<td>0.06</td>
<td>0.09</td>
<td>0.13</td>
<td>0.07</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Path to Success: Mobility of the professors
Table 6 shows the mobility of the professors inbetween the states. We see that some states such as California is the human capital source of some other states such as Texas. 21 percent of the Texas professors get their PhD from California whereas only 2 percent of the California professors get their PhD degrees from Texas. The same type of human capital transfer is also true for undergraduate education but to a lesser degree. Like the PhD education, only 2 percent of the professors in California get their undergraduate degrees from Texas. However, only 12 percent of the Texas professors get their undergraduate degrees from California. Only 16 percent of the professors work in the states that they get their undergraduate degrees. This rate is much lower than the average graduate degree holder. Kodrzycki (2001) uses the National Longitudinal Survey of Youth which is a nationally representative data-set. There are 6000 observations who are 14 to 22 years old in 1979. Kodrzycki (2001) notes that only 40 percent of the sample moved to another state after college graduation when the sample is surveyed again in 1996. This is a much lower mobility than the mobility observed from our sample.

Table 6. Mobility of professors inbetween states

<table>
<thead>
<tr>
<th>From</th>
<th>CA</th>
<th>IL</th>
<th>MA</th>
<th>MI</th>
<th>NJ</th>
<th>NY</th>
<th>PA</th>
<th>TX</th>
<th>WA</th>
<th>WI</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>0.28</td>
<td>0.11</td>
<td>0.09</td>
<td>0.12</td>
<td>0.12</td>
<td>0.11</td>
<td>0.12</td>
<td>0.12</td>
<td>0.20</td>
<td>0.10</td>
</tr>
<tr>
<td>IL</td>
<td>0.04</td>
<td>0.11</td>
<td>0.04</td>
<td>0.07</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.05</td>
<td>0.04</td>
<td>0.07</td>
</tr>
<tr>
<td>MA</td>
<td>0.14</td>
<td>0.16</td>
<td>0.33</td>
<td>0.10</td>
<td>0.18</td>
<td>0.21</td>
<td>0.13</td>
<td>0.09</td>
<td>0.11</td>
<td>0.13</td>
</tr>
<tr>
<td>MI</td>
<td>0.03</td>
<td>0.05</td>
<td>0.02</td>
<td>0.14</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>NJ</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.02</td>
<td>0.09</td>
<td>0.05</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>Other States</td>
<td>0.28</td>
<td>0.28</td>
<td>0.24</td>
<td>0.38</td>
<td>0.24</td>
<td>0.27</td>
<td>0.31</td>
<td>0.36</td>
<td>0.26</td>
<td>0.30</td>
</tr>
<tr>
<td>Total</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Path to Success: Mobility of the professors
Table 6 shows the mobility of the professors inbetween the states. We see that some states such as California is the human capital source of some other states such as Texas. 21 percent of the Texas professors get their PhD from California whereas only 2 percent of the California professors get their PhD degrees from Texas. The same type of human capital transfer is also true for undergraduate education but to a lesser degree. Like the PhD education, only 2 percent of the professors in California get their undergraduate degrees from Texas. However, only 12 percent of the Texas professors get their undergraduate degrees from California. Only 16 percent of the professors work in the states that they get their undergraduate degrees. This rate is much lower than the average graduate degree holder. Kodrzycki (2001) uses the National Longitudinal Survey of Youth which is a nationally representative data-set. There are 6000 observations who are 14 to 22 years old in 1979. Kodrzycki (2001) notes that only 40 percent of the sample moved to another state after college graduation when the sample is surveyed again in 1996. This is a much lower mobility than the mobility observed from our sample.
Table 7 shows how the mobility changes through time. Due to the improvements in internet and transportation technologies it is natural that the professors are more likely to travel to greater distances for their education and work. However, we see from the second column of Table 7 that the distance between undergraduate institution and the current university of the professor actually decreased by 2 miles between 1970s and 1990s. The fourth column shows that the mobility between the graduate degree and the current university of the professor just changed around 50 miles from 1970s to 1990s. The largest mobility increase is between the undergraduate and graduate degrees. The students move around 200 miles more to get their graduate degrees. The bottom line is that the professors always moved for their jobs. The distance did not matter much. However, they are more mobile for their PhD degrees recently.

### Table 7. Mobility of Professors Change Through Time (miles)

<table>
<thead>
<tr>
<th>PhD Degree Attained</th>
<th>UG-Current</th>
<th>UG-G</th>
<th>G-Current</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before 1970</td>
<td>845</td>
<td>643</td>
<td>889</td>
</tr>
<tr>
<td>1971-1980</td>
<td>940</td>
<td>728</td>
<td>947</td>
</tr>
<tr>
<td>1981-1990</td>
<td>933</td>
<td>846</td>
<td>954</td>
</tr>
<tr>
<td>1991-2000</td>
<td>938</td>
<td>922</td>
<td>1003</td>
</tr>
<tr>
<td>After 2000</td>
<td>963</td>
<td>943</td>
<td>1004</td>
</tr>
<tr>
<td>All</td>
<td>928</td>
<td>820</td>
<td>964</td>
</tr>
</tbody>
</table>

**Path to Success: Inbreeding**

The degree of inbreeding is measured in Table 8. The number of professors who work in the same university that they get their undergraduate (PhD) degrees is divided to the total number of professors in the second (third) column. Table 8 shows that the ratio of professors who work in the same university that they get their degrees is 4 percent. The ratios are very similar for both undergraduate and graduate degrees. The universities which have high inbreeding from undergraduate degrees have also high inbreeding for the PhD degrees. The correlation between the two ratios is 0.89. Only universities that the ratio exceeds 10 percent is Harvard and MIT. The inbreeding rate for PhD degrees is around 7 percent for Princeton which has top PhD programs. The most
surprising result is University of Chicago which is another top university. The inbreeding rate for the undergraduate degrees is just 0.5 percent and PhD degrees is just 2.7 percent.

Table 8. The degree of inbreeding

<table>
<thead>
<tr>
<th>UG/Prof</th>
<th>G/Prof</th>
<th>UG/Prof</th>
<th>G/Prof</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIT</td>
<td>0.146</td>
<td>0.162</td>
<td>Minnesota</td>
</tr>
<tr>
<td>Harvard</td>
<td>0.145</td>
<td>0.179</td>
<td>Chapel Hill</td>
</tr>
<tr>
<td>UT Austin</td>
<td>0.074</td>
<td>0.050</td>
<td>Brown</td>
</tr>
<tr>
<td>Yale</td>
<td>0.067</td>
<td>0.079</td>
<td>UCLA</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>0.066</td>
<td>0.027</td>
<td>UC Santa Cruz</td>
</tr>
<tr>
<td>Rice</td>
<td>0.064</td>
<td>0.064</td>
<td>TAMU</td>
</tr>
<tr>
<td>Caltech</td>
<td>0.063</td>
<td>0.101</td>
<td>UC San Diego</td>
</tr>
<tr>
<td>Cornell</td>
<td>0.061</td>
<td>0.065</td>
<td>Northwestern</td>
</tr>
<tr>
<td>Berkeley</td>
<td>0.060</td>
<td>0.022</td>
<td>Rutgers</td>
</tr>
<tr>
<td>Princeton</td>
<td>0.053</td>
<td>0.068</td>
<td>Univ Pittsburgh</td>
</tr>
<tr>
<td>Washington U</td>
<td>0.050</td>
<td>0.019</td>
<td>Colorado</td>
</tr>
<tr>
<td>Univ Florida</td>
<td>0.050</td>
<td>0.032</td>
<td>UC Davis</td>
</tr>
<tr>
<td>Michigan S.</td>
<td>0.046</td>
<td>0.030</td>
<td>Univ Arizona</td>
</tr>
<tr>
<td>Michigan</td>
<td>0.046</td>
<td>0.026</td>
<td>NYU</td>
</tr>
<tr>
<td>Univ Utah</td>
<td>0.044</td>
<td>0.029</td>
<td>Johns Hopkins</td>
</tr>
<tr>
<td>Upenn</td>
<td>0.043</td>
<td>0.043</td>
<td>UC Irvine</td>
</tr>
<tr>
<td>Duke</td>
<td>0.043</td>
<td>0.035</td>
<td>Arizona State</td>
</tr>
<tr>
<td>Purdue</td>
<td>0.043</td>
<td>0.026</td>
<td>Maryland</td>
</tr>
<tr>
<td>Penn State</td>
<td>0.043</td>
<td>0.048</td>
<td>Ohio State</td>
</tr>
<tr>
<td>U. Washington</td>
<td>0.042</td>
<td>0.027</td>
<td>Vanderbilt</td>
</tr>
<tr>
<td>Urbana</td>
<td>0.042</td>
<td>0.045</td>
<td>Boston Univ</td>
</tr>
<tr>
<td>Carnegie M.</td>
<td>0.041</td>
<td>0.045</td>
<td>South Calif</td>
</tr>
<tr>
<td>Columbia</td>
<td>0.041</td>
<td>0.044</td>
<td>Chicago</td>
</tr>
<tr>
<td>Stanford</td>
<td>0.039</td>
<td>0.060</td>
<td>UC Santa Barbara</td>
</tr>
<tr>
<td>All 48 Univ</td>
<td>0.041</td>
<td>0.039</td>
<td></td>
</tr>
</tbody>
</table>

Path to Success: Time to Graduate
The students rush to complete their PhD degrees because the prospective employers perceives a negative signal for late graduates. We see that this perception has a statistical basis. Table 9 compares the completion rate of the professors to the average PhD holders. We see that the median elite academicians complete the PhD degree much earlier than the median PhD holder. For instance, the median engineering professor graduate in six after she gets her undergraduate degree in 1990s whereas the median engineering degree takes 2.8 years more in 1993.

Table 9. Median time lapse between undergraduate and PhD degrees by academic fiels.

<table>
<thead>
<tr>
<th>From Thurgood and Clarke (1995)</th>
<th>Natural Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Sciences</td>
<td>5</td>
</tr>
</tbody>
</table>
Conclusion
Many academic dreams start at the early childhood. The students who have an academic tendency may want to know about their chances to be a professor at some prestigious university at an early stage. The previous studies focus on the success of PhD alumni in the academic world. We take one step backward and show the chances of undergraduate alumni to become professors. We show that undergraduate degrees attained by elite academics are more dispersed than their PhD degrees. Therefore, there is more opportunity for the graduates of lower ranked universities at the undergraduate level to become professors in a prestigious universities than the graduates of the lower ranked universities at the PhD level.
The students who aim to become professors at some prestigious university should expect to be much more mobile than their classmates. If the students dream to work in the same university that they attend then they will be disappointed to learn that only 4 percent of the professors work in the same university that they graduate. The students should listen to their advisors who tell them to finish their thesis on time. The PhD graduates who become professors in prestigious universities graduate much faster than the median PhD graduate.
The low degree of inbreeding implies that many universities employ the academics trained from other universities. The high ranking institutions provide human capital for the lower ranking institutions. Some states such as California provide human capital for other states such as Texas. Such human capital transfers between states show the brain drain dynamics within United States. Private universities provide human capital for public universities. This is an important justification for public support for private universities.

References
The list of these universities in alphabetic order are: Arizona State, Boston Univ, Brown, Caltech, Carnegie Mellon, Columbia, Cornell, Duke, Harvard, Johns Hopkins, Michigan State, MIT, Northwestern, NYU, Ohio State, Penn State, Princeton, Purdue, Rice, Rutgers, Stanford, Texas A&M, UC Berkeley, UC Davis, UC Irvine, UC Los Angeles, UC San Diego, UC Santa Barbara, UC Santa Cruz, UNC Chapel Hill, Univ Arizona, Univ Chicago, Univ Colorado, Univ Florida, Univ Illinois-Urbana, Univ Maryland, Univ Michigan, Univ Minnesota, Univ Pittsburgh, Univ South Calif, Univ Utah, Univ Washington, Univ Wisconsin, Upenn, UT Austin, Vanderbilt, Washington U, Yale

The excluded universities are: Rockefeller University, UC San Francisco, UT Southwestern Medical Center
Abstract
This study explores and analyzes three types of potential knowledge in international scientometrics from weak ties, indirect reference, and the ISI-S model. To explore the flow of potential knowledge, journals dual-map overlays technology is applied. Such an approach is consistent with the weak ties, thereby broadening the perspective of research on potential knowledge.

Conference Topic
Knowledge discovery and data mining

Introduction
Potential knowledge is a concept that corresponds to mainstream knowledge, the scope of which includes undisclosed public knowledge except for mainstream knowledge (Donald R. Swanson, 2000). The study of the flow of potential knowledge is a necessary part of improving the knowledge system. Potential knowledge is not yet discovered; however, it may be subversive. The development of knowledge is also a process wherein potential knowledge and mainstream knowledge alternate. Much hidden knowledge is found in the realm of science metrology which cannot be perceived by people, because it is outside mainstream knowledge. The analysis of the type of potential knowledge and the revelation of its flow direction contribute to the development of scientific metrology.

Potential knowledge exists in many ways and is mainly derived from published and unpublished papers. The former is chiefly implied in weak ties with indirect references; the latter cannot be calculated because of its being unpublished, but it can be analyzed by the ISI-S model.
1 Data and methods

As a relatively sound knowledge system, scientific metrology is chiefly constructed by mainstream knowledge. The purpose of this research is to emphasize and reveal the value of potential knowledge via statistics and analysis of literature related to international scientific metrology in the past five years.

Using the “advanced search” function of the Thomson Reuters Web of Science TM (WOS) from the library of Wuhan University, relevant data on “TS = Scientometrics or TS = Bibliometrics or TS = Informetrics” for the years between 2012 and 2016 are collected. Subsequently, 2341 papers were found on scientific metrology as of 1st January 2016.

In this paper, we analyze these data quantitatively and qualitatively using the methods of bibliometrics, content analysis, knowledge maps, and social networks.

2 Analysis of potential knowledge in weak ties

Weak ties theory is proposed by the sociologist Grannovetter. He argues\(^1\)\(^2\) that weak ties represent a transient social contact between two actors; therefore, it is less structured and can connect different groups or subgroups and deliver messages to different parts of network. Doreian (1979:51-52) reveals the characteristics of cliques. If only relationships are considered, without the direction of the lines, and if all connections are considered as reciprocal relationships, then the results of the analysis reveal weak cliques, which represent a form of indirect relations.

According to Zipf’s Law, in the low frequency words, the graph of the network knowledge can be drawn in low frequency words, thereby indicating that small group network cliques is the key content among low frequency words, suggesting latent knowledge in weak ties in international scientific metrology. A prevailing trend states that if increasing number of researchers aim at the potential knowledge, it could become mainstream knowledge in the near future. Figure 1 is the map of low frequency key words in international scientific metrology over the past five years, in which the box represents the keyword. The larger the boxes are, the higher the frequency of the key word, and the lines between the boxes represent the relationships between them. Three large groups of high frequency words are found in the maps: 1. Quantity, Performance, Quasity, Normalization, Energy, etc.; 2. Neural regeneration, nerve repair, nerve transfer, nerve regeneration, etc.; and 3. University ranking, research competitiveness, research ranking, Google Maps, spatial scientometrics, etc. All of these groups are valuable latent knowledge in weak ties; thus, we should focus on the mainstream of future knowledge.
The analysis of conventional measurement usually focuses on hot spots, that is, high frequency words are taken as the study object. However, low frequency words often connect different groups or subgroups and deliver messages to the network. Their specific features are described as extensive, heterogeneous, and intermediary. Low frequency words within the weak ties are consistently free from networks. Given their low frequency, with few co-occurrence words and formation of network nodes, they can hardly form group factions or they can only form small scape networks. However, once the association exists, it becomes latent knowledge.

3. Analysis of latent knowledge in indirect reference associations

The journals dual-map overlays technique is a new tool to reflect the flow of knowledge, based on the coupling and co-citation analysis of journals. Bibliographic coupling, co-citation, and journals dual-map overlays can be used to explore some of the internal knowledge correlations of international scientific metrology research. They can also analyze the flow of knowledge between journals composed of citations and the citations of the 2,341 international journals of scientific metrology.

3.1 Coupling of international journals of scientific metrology

Journal coupling analysis establishes the relationships among journals by evaluating the degree of coincidence of different journals that have the same cited journals. The greater the number of documents coupled by the two journals, the more similar those two are. The co-citation of journals refers to the number of papers that have been published in two journals; and the greater the number of co-citations, the more similar the two journals. Coupling analysis of journals was conducted via VOSviewer³, developed by the Center for Science and Technology Studies at the University of Leiden, the Netherlands. The VOS Mapping method is applied to determine the coordinates of the elements in the two-dimensional space.⁴ To identify the distribution
of the most important journals, we use the density maps to visualize the journals. (In the density map, the greater the weight of the nodes, the more journals around them, and the greater the density around them.) The total cited networks and the journals dual-map overlays are obtained, and we subsequently use CiteSpace (developed by Prof. Chen) to analyze them.

Figure 2 shows a density chart of the coupling analysis of international scientific metrology research journals in the recent five years. Font size characterizes the amount of papers published in different journals such that the former is directly proportional to the latter. Network density is an important indicator of the overall level of network cohesion; the greater the network density, the more full the knowledge and information exchange. Distance between the points in two-dimensional space reflects the similarity between journals. Hence, the closer they are, the greater their similarity. Through the density map, journals can be clearly divided into a scientific metrology-based large community and several other small communities, by which we can know the tendency of achievement in international scientific metrology. Further analysis indicates that international scientific metrology represented the best in library and information science such as scientometrics, Journals of informetrics, and others, with full load of knowledge and communication. These results are consistent with the normal development track in this field, and reflect that much room for development remains for other disciplines.

3.2 Cited journals in international scientific metrology
Cited journals in international scientific metrology refer to journals that are cited in scientific econometrics studies, that is, the journals published by scholars of international scientific metrology as a reference list of articles published in the journal itself. Citation can reflect the knowledge flow from the cited journal to citing journals, and the cited journal can be considered the knowledge provider of international scientific metrology or the knowledge base. We analyzed the total citations in international scientific metrology literature, as shown in Figure 3, which is similar to
the density map. A sizeable scientific metrology-based group is also observed. As shown in Bibexcel statistics, for the top 20 journals with the achievement of scientific metrology and citing journals, all are in the realm of scientific metrology or library information science. Furthermore, many journals occur in both the citing journals and the cited journals, thereby indicating that the knowledge flow is concentrated in these journals. Such journals represent an important knowledge base of international scientific metrology.

![Figure 3 Analysis of co-citations of international scientific metrology over the past five years](image)

3.3 Potential knowledge flow in journals dual-map overlays

The development of information visualization has increased interest in issues related to knowledge tracking. Information visualization can help us illustrate the dynamic changes in the scientific paradigm as well as explain how to accommodate potential knowledge and mainstream knowledge in one visible framework. The distribution of potential knowledge flow in scientific metrology can be analyzed by the journals dual-map overlays technique in Citespace V. Journal dual-map overlays, a technique developed by Professor Chen and Leot Leydesdorff in 2013, provides journal similarity maps via computer visualization. Compared with the early journal overlay map, double graphs can provide more information and potential knowledge association, and show the knowledge flow among journals through the citation path of a paper.

Journals dual-map overlays is a novel method that shows the distribution, citation trajectory, and gravity drift of papers. The design of dual-map overlays describe citing overlay and cited overlay mapping in the same view, thereby showing an integrated citation instance and providing a clear origin and orientation at a glance. Such design also shows the possibilities in cross-subjects of both source and goal. Many overlays are added to basic dual-maps, and each layer must contain the reference record, such as records obtained from Web of Science searches. While the smallest set may contain only a single paper, no limit exists for the peak set. In the journals dual-map overlays, each reference instance is represented by an arc. Such arc originates from the source journal of the citing paper, and the end of the arc is the target journal in the cited base map. Arcs from the same set are represented with the same color to differentiate the
different sets’ cite patterns by their unique color. Base maps provided by panoramic scientific maps enable the analysis of the interaction map. As citing journals and cited journals are shown simultaneously in dual-map overlays, comparing reference behavior of different groups via source journals and target journals is possible.\textsuperscript{10}

Journals cited in international scientific metrology also reveal that knowledge flows from citing journals to cited journals. Thus, international scientific journals can be divided into citing journals and cited journals. Figure 4 is a dual-map overlays of citing journals and cited journals in international scientific metrology over the past five years. The chart on the left is a journal cluster chart based on the citing journal map, whereas that on the right is based on the cited journal map.\textsuperscript{11} Figure 4 shows that both the source journals (on the right) and target journals (on the left) have two groups. Source journals flow from molecular, biology, genetics, and nursing, to medical, neurology and movement. Conversely, target journals flow from molecular, biology, genetics, nursing, medical, psychology, and pedagogy to mathematics, cultivation, ecology, and zoology. This outcome reveals the latent knowledge flow in international scientific metrology, and is consistent with the weak ties in neurology in indirect relations. Concepts in neurology include the following: neural regeneration, nerve repair, nerve transfer, and nerve regeneration. Moreover, the consistency reflects the regularity of mining potential knowledge, thereby confirming the value of this paper.

Figure 4 Journals dual-map overlays in international scientific metrology from 2012 to 2016

4 Potential knowledge of ISI-S model
According to Merton (1973), the goal of scientific systematization is to extend the knowledge that has been identified. This systematic approach prevents the omission of any content that is not easily detected and promotes the improvement of the knowledge system. The scientific metrology model (ISI-S model) designed by Vinkler
(2002a) is a global network which considers information system and scientific knowledge as a whole, with close connections, and with dynamically changing network content and scale. The contents and rules of a single cluster are regulated by the relevant assessment process, and this model is used to explore the changes of potential knowledge in unpublished papers, thereby providing the basis for the scientific system of international scientific metrology. The ISI-S model builds six main sets of information: information in published articles, aged or neglected information, information with short-term effects, information with long-term effects, basic scientific knowledge, and scientific information. Each information set contains potential information that can complement the potential knowledge of international scientific metrology from a static perspective. The level of information set represents the hierarchical level of the following scientific system (see Figure 5): 

Figure 5 Metrological model of scientific metrology information (ISI-S model)
Note: The map is from The Evaluation of Research by Scientometric Indicators

As shown in Figure 5, each information set has a relationship with potential knowledge, such as “information to be modified or rejected in literature” in literature to be published, “information that has not been assessed or lost after assessment” in publications, and “unrelated (wrong or redundant) information and information with a hidden impact” that may be absorbed and evaluated. Other knowledge includes “information with short-term effect,” “information with long-term effect,” “aging information,” “aging knowledge,” and “wrong or superfluous knowledge” in basic scientific knowledge or scientific knowledge. They may be treated as wrong, redundant, and aged information, and will then be merged or discarded. However, such knowledge may possess hidden value to be excavated or transformed into information with significant effects in the future. Merged or discarded information
may form basic scientific knowledge over a long period. Fundamental scientific knowledge is composed of pieces of information which are effective in the long term, representing a combined, systematized, widely accepted information system of one unit (as a field or subject). The resulting information system may have a significant effect on each subject, research area, or on scientific research. Based on the ISI-S model, publications that have not been cited for a long time can be considered aging, unrelated information with latent influence. Thus, much scientific research is repetitive work that produces many publications that are unaffected or less influential. Therefore, in every information integration phase, potential influence knowledge, irrelevant knowledge, and aging knowledge exist. This potential knowledge plays an essential role for the integrity of the knowledge system by improving the knowledge from the perspective of unpublished literature.

5 Conclusion
In conclusion, many ways exist to develop and exploit the latent knowledge of international scientific metrology. In this research, we analyze the potential knowledge in weak ties, the latent knowledge in indirect reference, and the latent knowledge in the ISI-S model from published and unpublished literature. We also conclude that potential knowledge in weak ties is consistent with the target knowledge in the journals dual-map overlays of knowledge flow, thereby validating the hypothesis of this paper.

Similar to most disciplines, international scientific metrology is a relatively self-sufficient discipline that consists of mainstream knowledge in a system, thereby keeping many scholars engaged in the long term. However, as discussed, international scientific metrology contains considerable potential knowledge with the possibility of being activated as mainstream knowledge. Subsequently, such knowledge becomes basic scientific knowledge or common science, thereby promoting the development of disciplines. For example, the number of occurrences of the keyword “altometrics” during the year 2012–2016 are 0, 2, 6, 18, and 13 (as analyzed by Bibexel software). Evidently, this keyword presents itself as a kind of potential knowledge in weak ties. In fact, the prevailing trend suggests that altometrics is becoming a research hot spot in international scientific metrology.

References
Van Eck NJ, Waltman L, Dekker R, van den Berg J. A comparison of two techniques for


Chen C. Mapping Scientific Frontiers: The Quest for Knowledge Visualization: Springer; 2013.


Mapping the Semantic Word Shifts in Topics in the Field of Information Retrieval

Baitong Chen\textsuperscript{1} Ying Ding\textsuperscript{2} Feicheng Ma\textsuperscript{3}

\textsuperscript{1}baitongchen@shu.edu.cn
Shanghai University , Shanghai (China)

\textsuperscript{2}dingying@indiana.edu
Indiana University, Bloomington (America)

\textsuperscript{3}fchma@whu.edu.cn
Wuhan University, Wuhan (China)

Abstract
Understanding semantic word shifts in the scientific domains is essential for facilitating interdisciplinary communication. Using a data set of published papers in the field of Information Retrieval (IR), this paper studies the semantic shifts of words in IR based on mining the per-word topic distribution over time. From our results, different patterns of word migrating among topics are recognized, and the semantic shifts represented by the migrating process are analysed accordingly.

Conference Topic
Knowledge discovery and data mining

Introduction
Interdisciplinary communication is a crucial issue for promoting innovation and science development. Today, data are extremely rich but overwhelming (Ding & Stirling, 2016), different research areas are developing new ideas and methods constantly. During the evolution of scientific topics, different topics may view same words or phrases in different ways, e.g., “data science” in biomedical science mainly means gene codes or protein structure, and in computer science it focus more on algorithm designing, while in information science it cares more about people and society.

The phenomena of the same word appear in multiple topics are essentially semantic shifts of words during the topic evolution process. In practice, a word shifts or extends its semantics while a topic evolves, where the same word can appear in multiple topics with different contexts. In data mining and computational linguistic areas, language processing tools has been developed to recognize semantic word shifts. However, existing work studies the semantic shifts of words mainly to facilitate natural language processing in information retrieval problems, where the relationship of words with topics has been overlooked.

This paper studies semantic word shifts in different topics in a scientific domain, and using a data set in the field of information retrieval as an empirical study. A word’s semantic shifts via topic channels is herein defined by its related innovations and applications. When the same word distributed in different topics, even if it is embedded in a stable context (namely, the linguistic semantics based on the distributional hypothesis), it is usually in associate with different innovations and applications due to the change of topic themes.
Related work

Topics are essentially collections of words with semantic functions (Griffiths, Steyvers, Blei, & Tenenbaum, 2005). Defined as a change of one or more meanings of the word in time (Lehmann, 1993), the shifts of word semantics and its detection has been the focus of much research in recent years. Studies on word semantics over time can be viewed from two perspectives: synonymy detection and polysemy detection. Synonymy detection monitors the use of different words with the same meaning over time (Kenter, Wevers, Huijnen, & de Rijke, 2015). Based on a small set of input words in a certain time period, ranked lists of terms for a consecutive series of periods in time would be output. The words in the ranked lists are meant to denote the same concept as the input words. Studied more extensively, polysemy detection monitors different meanings expressed by the same word over time. A word can change semantically in a way wherein new meanings replace the old ones, or acquire additional meaning with the original meaning may still be widely used (Wijaya & Yeniterzi, 2011).

In polysemy detection, distributional semantic models (Gulordava & Baroni, 2011; Hamilton, Leskovec, & Jurafsky, 2016; Kim, Chiu, Hanaki, Hegde, & Petrov, 2014) are widely used for quantitative measurement. In these models, the similarity between words is measured by vector space models where each word is associated with its context vectors. Existing works have studied the semantic shifts of words mainly to facilitate natural language applications, for example, time-aware query expansion for document retrieval tasks in a historical corpus. In this case, words are studied alone and are not associated with topics.

Methods

Data set and topic extraction

Information retrieval (IR) is chosen as the target domain. Papers are collected from Web of Science for 1956-2014, making a total of 20,359 documents, with search based on a set of IR-related terms. Search term selection refers to the paper by Xu et al. (2015). The selected document types include article, book, book chapter and proceedings paper. The title and abstract fields are used as the text corpus for extracting topics. Before extracting topics, all terms are stemmed using the Porter2 stemming algorithm. A stop word list (Yan, Ding, Milojević, & Sugimoto, 2012) is used to filter common words. Words with only one letter or appear less than five times are removed.

The Latent Dirichlet Allocation (LDA) proposed by Blei et al. (2003) is applied for extracting topics from the corpus. LDA is a three-layer Bayesian model that is now widely used in discovering the latent topic themes in collections of documents. The LDA model represents each document with a probability distribution over topics, where each topic is represented as a probability distribution over words. For a detailed explanation of the algorithm, refer to, e.g., Blei (2012). The Gensim library (Rehurek & Sojka, 2010) is used for implementing the LDA model, where the parameters are set as the standard value proposed by Gensim. Considering the size of the dataset and the experience from previous studies (Ding, 2011), the number of topics is set at five.

Mapping word semantics via topic channels

We define the same word appears in different topics as word migration, which resemble the migration of populations in demographics as if the word is an analogue of a human population and topics are territories.
A word’s migration is examined herein based on its yearly topic distribution. The topic distribution of a word in a document is parameterized by the variational parameter $\phi$ in the LDA model (Hoffman, Bach, & Blei, 2010). The $\phi$ value indicates the likelihood of a word belonging to a topic in terms of a particular document. Each word in a document has five topic $\phi$ values corresponding to the five global topics (Table 1). After normalizing by the frequency of the word in the document, the sum of a word’s $\phi$ values is equal to 1. The same word from two different documents usually has two different sets of $\phi$ values. For a selected word, we calculate its average $\phi$ value for each topic in each year. The average $\phi$ values represent the topic probability distribution for the word in that year. As the $\phi$ values change over the years, the word migrates between topics over time.

Table 1. An example of the normalized $\phi$ values of a word in a particular document

<table>
<thead>
<tr>
<th>Word</th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.45</td>
<td>0.05</td>
<td>0.12</td>
<td>0.21</td>
<td>0.17</td>
<td>1</td>
</tr>
</tbody>
</table>

Results and discussions

**Topic extraction**

The top ten words with the highest probabilities of the five extracted topics are presented in Table 2. Topic 1 focuses on user-oriented problems, covering online information-seeking behavior, use of digital resources such as digital library by research scholars, and user information needs, especially for health information search. The main theme of topic 2 centers on multimedia information retrieval, especially image retrieval. Topic 3 and topic 4 both study queries for structured data sets, where topic 3 mainly deals with traditional query processing for relational and object-oriented databases, and topic 4 primarily focuses on distributed query processing for spatial networks and communication networks. Topic 5 studies text retrieval for unstructured documents, which involves document indexing and terminology processing problems, such as term disambiguation, query expansion, and cross-language retrieval.

Table 2. Top words in the topics

<table>
<thead>
<tr>
<th>word</th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>research</td>
<td>image</td>
<td>query</td>
<td>query</td>
<td>document</td>
<td></td>
</tr>
<tr>
<td>user</td>
<td>feature</td>
<td>data</td>
<td>data</td>
<td>text</td>
<td></td>
</tr>
<tr>
<td>data</td>
<td>content</td>
<td>database</td>
<td>network</td>
<td>user</td>
<td></td>
</tr>
<tr>
<td>design</td>
<td>similarity</td>
<td>language</td>
<td>algorithm</td>
<td>relevance</td>
<td></td>
</tr>
<tr>
<td>library</td>
<td>visual</td>
<td>relational</td>
<td>time</td>
<td>term</td>
<td></td>
</tr>
<tr>
<td>web</td>
<td>music</td>
<td>semantic</td>
<td>index</td>
<td>query</td>
<td></td>
</tr>
<tr>
<td>find</td>
<td>learn</td>
<td>integration</td>
<td>distributed</td>
<td>web</td>
<td></td>
</tr>
<tr>
<td>analysis</td>
<td>algorithm</td>
<td>structure</td>
<td>optimization</td>
<td>evaluation</td>
<td></td>
</tr>
<tr>
<td>medical</td>
<td>object</td>
<td>computing</td>
<td>rank</td>
<td></td>
<td></td>
</tr>
<tr>
<td>access</td>
<td>color</td>
<td>knowledge</td>
<td>tree</td>
<td>word</td>
<td></td>
</tr>
</tbody>
</table>

**Topic distribution**

Figure 1 presents an overview of the topic distribution status of all words. Each node is a word, and is represented by a five-dimensional vector, which is equivalent to the per-word topic distribution over the whole corpus. Words with similar topic distribution are closer to each other, and words with dissimilar topic distribution are far apart from each other, which result in the clearly distinguished five clusters corresponding to the five topics. Words that are stable in
one topic are far away from the center, such as Color and Image in Image Retrieval. Words that are distributed about evenly in multiple topics are closer to the center, such as Web, Context and Evaluation. Words that migrates between two topics are correspondingly located between these two topics, e.g., Language between Database Querying and Text Retrieval.

Figure 1. Per-word topic distribution

Semantic word shifts
Figure 2 shows the migration of the top 10 words with the highest probabilities in each topic over time. Because topics share some of the top words, there are 42 unique top words in total.
Considering the number of topics between which a word mainly migrates, words can be categorized into three groups: the non-migration, the dual-migration, and the multi-migration. Non-migration words are those that always belong to only one topic over time. In Figure 2, non-migration words include, for example, Color and Image in Image Retrieval; Document, Text and Word in Text Retrieval; Library in User Study; Network and Tree in Query Processing. Words belong to the non-migration type are core words in the topic, which reflect the research theme or main method of a topic. Dual-migration words are distinguishable in two topics. A typical example is Language in Database Querying and Text Retrieval. When in Database Querying, it is related with query language in databases. When in Text Retrieval, it is placed in the situation of cross-language retrieval. Multi-migration words are those migrates among multiple (more than two) topics, most words belong to this type. There are multiple underlying ideas represented by these words.

Considering the stability of the probabilities of a word in different topics over time, there are two types needed to be noted: the divergent words and the convergent words. Divergent words start with an approximately even probability in several topics, but the probabilities diverge in later periods for the word being assigned more apparently to one or two topics, e.g., Web starts with a probability of 0.3 distributed evenly in User Study, Database Querying and Text Retrieval. The probability in Text Retrieval then increases and separates from the other two, as web is studied more specifically under the context of text retrieval in web environments in later periods. The diverging process of a migrating word indicates the word is becoming topic specific. The topic it mainly migrates towards becomes increasingly developed, resulting in the word being more studied under the specific context of this topic.
Convergent words are obviously assigned to one topic at first, but the probabilities in the topics converge in later periods, resulting in the word being distributed evenly in multiple topics. Taking semantic as an example (Figure 2), semantic is assigned mostly to Database Querying appearing as semantic query language at first, then the word gradually migrates to Text Retrieval and Image Retrieval, and the probabilities in the above three topics converge in the 2010s. The converging process of a migrating word indicates the increasing importance of the word in the domain, for it is not only studied in a specific topic, but developed various applications covering broader range of contexts in different topics. In the case of semantic, it is not only studied as semantic query language as in the beginning, but also studied as semantic text retrieval in Text Retrieval and image semantics in Image Retrieval in later periods.

Conclusion
This paper studied the semantic word shifts of topics in the field of information retrieval. Based on the per-word topic distribution over time, different patterns of word migrating among topics are recognized. Our work contributes to a better understanding of interdisciplinary communication, and further direction could consider mining the local context shifts during the semantic shifts.

Acknowledgments
This work is funded by the National Natural Science Foundation of China (71420107026).

References
https://doi.org/10.1145/2064448.2064475
Xu, J., Ding, Y., & Malic, V. (2015). Author Credit for Transdisciplinary Collaboration. PLOS ONE, 10(9), e0137968. https://doi.org/10.1371/journal.pone.0137968
Word semantic change: The law of differentiation vs. the law of parallel change

Erjia Yan¹ Yongjun Zhu²

¹ey86@drexel.edu
Drexel University, Philadelphia (U.S.A.)

²yoz2008@med.cornell.edu
Weill Cornell Medicine, Cornell University, New York (U.S.A.)

Abstract
The objective of this research-in-progress paper is to reveal word semantic change and verify whether the change conforms to the law of differentiation or the law of parallel change. This paper identifies a set of representative words in biomedical literature based on word frequency and word-topic probability distributions. It employs a skip-gram word2vec model to the identified words to measure word-level semantic changes. This paper finds no overwhelming evidence to support either the law of differentiation or the law of parallel change.

Conference Topic
Knowledge discovery and data mining; Methods and techniques

Introduction
Word semantic change is one of the several forms of historical linguistics changes that include sound change, grammatical and syntactic change, and semantic change (Xu & Kemp, 2015). Among them, semantic change is the least understood (Crowley & Bowern, 2010). A few linguistic laws have been proposed to study semantic change (Hamilton, Leskovec, & Jurafsky, 2016b; Xu & Kemp, 2015) such as the law of differentiation that dictates that synonyms tend to differentiate in meaning over time and the law of parallel change that governs that related words tend to undergo parallel changes. These laws are rooted in classic linguistic literature and, recently, two advances have made it possible to examine word semantic change and verify these laws with large empirical data. First, the availability of large-scale collections of diachronical textual data has markedly facilitated scholars’ computational investigations of language change (Frermann & Lapata, 2016). Projects such as Google Books (Lin et al., 2012) and Corpus of Historical American English (Davies, 2010) have an extensive coverage dating back to the early 1800s. Commercial databases such as Web of Science have also made an effort to index their collections to as early as the 1900s (Larivière, Archambault, & Gingras, 2008; Larivière, Gingras, & Archambault, 2009). Second, the advanced computational methods developed in the last few years for the first time have had the ability to harness large dynamic data.

A number of methods were developed and applied to understand word semantic change, including several word embedding algorithms such as positive point-wise mutual information (PPMI), singular value decomposition (SVD), and skip-gram models. Among these methods, word2vec has gained popularity in understanding word embeddedness and word semantic change. Word2vec is a distributed word representation technique from deep learning approaches (Manning, 2016). It was proposed by Mikolov and colleagues (2013) to estimate continuous vector representations of words from large corpora. It was built on previous work on neural net language models (NNLM) and the advantage of word2vec is that it significantly reduced the computational complexity. Two word2vec models were proposed, one is the continuous bag-of-words (CBOW) model that does not consider word orders and the other is...
the continuous skip-gram model that assigns different weights based on the proximity of words in a window.

The goal of this research-in-progress paper is to reveal word semantic change and verify whether empirical data support the law of differentiation or the law of parallel change by employing word2vec on a large data set of PubMed publications. The biomedical domain was selected primarily because the availability of large, publically available data sets in this domain that makes is possible to gain access to data and replicate research for transparency and sustainability.

Data

Abstracts of 18,777,129 articles published in the last 30 years (1987-2016) were downloaded from PubMed. The number of yearly publications ranges from approximately 360,000 for 1987 to more than 1,000,000 for the recent three years (2014-2016). To select a set of representative terms in the whole PubMed data set, we first identified top 500 words based on their frequencies in PubMed after a stop word list was applied. We then ran a topic model and identified, for each topic, 25 words with the highest probabilities. We merged the two lists and formed the final list of words for semantic distance calculation. Because a word can be in both the most frequent list and the highest probability list, the total number of words is 761.

Methods

In this section, we give a brief mathematical introduction to the algorithm behind the continuous skip-gram model using notations of Goldberg and Levy (2014). In this model, we consider the conditional probabilities \( p(c|w) \) given a corpus of words \( w \) and their contexts \( c \). The goal is to maximize the probability by setting the parameters \( \theta \) in \( p(c|w; \theta) \):

\[
\arg \max_{\theta} \prod_{(w,c) \in D} p(c|w; \theta)
\]

where \( D \) contains word and context pairs extracted from the corpus. \( p(c|w; \theta) \) can be parametrized using the softmax function such that \( p(c|w; \theta) = \frac{e^{v_c^T v_w}}{\sum_{c'} e^{v_{c'}^T v_w}} \) where \( v_c \) and \( v_w \) are vector representations of \( c \) and \( w \). Chiu, Crichton, Korhonen, and Pyysalo (2016) presented settings for each hyper-parameter that work best for PubMed publication data. Therefore, we used the parameter settings reported in their study. The parameters are set as follows: 10 (negative sample size), 1e-4 (sub-sampling), 5 (minimum-count), 0.05 (learning rate), 200 (vector dimension), and 30 (context window size).

We adopted the global measures by Hamilton, Leskovec, and Jurafsky (2016a). The global distance between a word in year \( t \) and year \( t+1 \) is calculated through: \( \cos(w_i^{(t)}, w_i^{(t+1)}) \), where \( w_i^{(t)} \) and \( w_i^{(t+1)} \) are word vectors of word \( w_i \) over all dimensions in \( t \) and \( t+1 \).

Results

To verify the law of differentiation and the law of parallel change, we identified a number of groups of synonymous words within the 761 words using an online thesaurus API. This API is developed based on source data from WordNet, the Carnegie Mellon Pronouncing Dictionary, and user suggestions. Because synonymous relations can be rather liberal in this API, two coders manually selected a more coherent set of synonymous words from the API’s output, resulting in 21 groups of synonymous words (87 words in total). Figure 1 shows the

1 https://words.bighugelabs.com/api.php
We inserted the average global distance of all 761 words as a benchmark (“Global Mean”).

We see in Figure 1 that synonymous words in most groups seemingly coevolve in the past 30 years. Nonetheless, when using the global mean as a benchmark, we found only two groups—group 8 (“assessment” and “evaluation”) and group 14 (“maximum”, “maximal”, and “peak”)—whose average correlation coefficient between the synonymous words is greater than the average correlation coefficient between the global mean and individual synonymous words of that group. Meanwhile, “medium” and “average” in group 5 and “genotype”, “gene”, and “dna” in group 11 have higher correlations with words in their respective groups than with the global mean, but when other synonymous words are added, as a group, words’ semantic change patterns relate more with the global mean than with each other. The results suggest that while synonymous words seem to coevolve semantically, their coevolving patterns are not significant when using the global mean as a benchmark. Thus, although the law of parallel change was identified in a few cases, this study did not find overwhelming evidence to support this law across the whole corpus.

As for the law of differentiation, among the 21 groups of synonymous words, two groups—group 9 (“experiments” and “research”) and group 17 (“therapeutic” and “curative”)—resulted in negative correlation coefficients between the synonymous words (-0.393 for group 9 and -0.392 for group 17). The majority groups, however, have strong, positive correlations between the synonymous words (average coefficient=0.65).

Conclusion
This research-in-progress paper did not find a consistent set of evidence to support the law of differentiation or the law of parallel change. PubMed is the single largest bibliographic data repository that is available to the public. We believe with the open data movement, there will be more high-quality publication data sets available to researchers. A future work in this direction will involve bringing in disciplinary perspective to gain insights on disciplinary epistemological differences invoked by word semantic changes.
Acknowledgments

This project was made possible in part by the Institute of Museum and Library Services (Grant Award Number: RE-07-15-0060-15), for the project titled “Building an entity-based research framework to enhance digital services on knowledge discovery and delivery”.

References


CS-LAS: A Scientific Literature Retrieval and Analysis System Based on Term Function Recognition (TFR)

Li Xin¹, Cheng Qikai² and Lu Wei³

¹lucian@whu.edu.cn
Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan(China)

²chengqikai0806@163.com
Information Retirval and Knowledge Mining Laboratory, Wuhan University, Wuhan(China)

³weilu@whu.edu.cn
Information Retirval and Knowledge Mining Laboratory, Wuhan University, Wuhan(China)

Abstract
We consider the problem of extant scientific literature search and analysis systems, and suggest that term function recognition can be very advantageous for fine-grain retrieval and semantic analysis of a huge amount of scientific literature in a specific domain. We first elaborate the definition of term function (TF) in the scientific literature context, and the status and development of term function recognition (TFR) globally. Then, the computer science domain is taken as an example to design and implement a system called CS-LAS, which can provide users with several useful function modules based on TFR, such as scientific information search and browsing, domain hot spot discovery, evolution trend detection, and scholarly recommendation based on correlation analysis of functional terms.

Conference Topic
Methods and techniques; Mapping and visualization; Knowledge discovery and data mining.

Introduction
Scientists can usually understand the knowledge structure and frontier of a research domain in which they are interested by reading relevant scientific literature, which is also a requisite link in the overall process of science. However, with the rapid growth of scientific literature and the increasing popularity of interdisciplinary collaborative research, it has become increasingly difficult for researchers to keep up with relevant developments only by manpower (Spangler et al., 2014). Therefore, it is of great importance for researchers to develop computer software to automatically and rapidly solve certain problems, such as how to quickly analyse a large amount of scientific literature to comprehensively and accurately identify the domain hot spots and evolution trends of a research field via computer technologies, how to automatically find a new research method for an existing research topic, and how to discover a research topic for which an existing method is appropriate. In past decades, scholars who are from the domains of information science, bibliometrics, digital library, and information visualization have discussed these issues, as well as a series of practical literature analysis systems, such as CiteSpace (Chen, Chen, Hou, & Liang, 2009), VosViewer (Van Eck & Waltman, 2009), NEViewer (Wang, Cheng, & Lu, 2014), and BICOMB (Cui, Liu, & Zhan, 2008).

Although these systems have played a critical and positive role in solving the above mentioned issues, they are usually based on term frequency, co-word clustering, and co-citation clustering analysis. In other words, they take the absolute occurrences of terms into account from the document level, and miss the semantic information contained by the terms in the scientific literature. When a term occurs in a scientific work, however, this does not mean that it constitutes the topic of the literature. Moreover, when a paper is cited, the cited part may not be the focal point of the article. Overall, simple quantitative statistical analysis of
literature at the document level only obtains surface information of terms, and cannot represent the deep semantic information contained in the terms, which frequently leads to inaccurate analytical results.

In fact, a term usually plays a specific semantic role in a specific research, and has a particular function that we refer to as term function (TF). The purpose of this study is to recognize the term function in relation to scientific literature and, on this basis, develop a literature analysis system called CS-LAS using data from the computer science field. CS-LAS can be effectively utilized for scientific literature retrieval, navigation, bibliometric analysis, and scholarly recommendation based on term function recognition.

The organization of this paper is as follows. In section 2, we review the relevant research on scientific literature analysis systems, term function (TF), and term function recognition (TFR) in scientific text. In section 3, we depict the design idea and construction of our scientific analysis system. In section 4, we elaborate the implementation of our system, including data collection, the construction of functional term sets, the implementation of retrieval and analysis modules, the design of the user interface, and the visualization of results. In section 5, we draw conclusions from this article and identify directions for future work.

Related research

Scientific literature analysis system (SLAS)

Scientific literature analysis system, which is absolutely critical in the era of big academic data, is an information analysis system that can automatically realize semantic analysis of a massive number of scientific papers based on theory and technology from statistics, information retrieval, and bibliometrics. According to the main function of the existing systems, we divided SLAS into two categories, one based on information retrieval and the other based on bibliometrics. The system based on information retrieval which uses database technology, mainly provides users with the function of literature retrieval, domain navigation and browsing, and simple statistical analysis of scientific papers. Traditional academic databases, such as Web of Science, Scopus, PubMed, etc., belong to this category. The system based on bibliometrics, such as CiteSpace (Chen, 2009), VOSViewer (Van, 2009), and NEViewer (Wang, 2009), can be used for co-citation analysis, co-word analysis, co-occurring analysis, and powerful visualization of scientific papers to help researchers quickly obtain knowledge of a research domain.

The two types of SLAS are currently widely used due to their major advantage of realization of automatically semantic analysis of scientific articles. However, much room for improvement of these systems exists to more fully satisfy users’ academic information needs. First, the system based on information retrieved mainly ranks the results by relevance, i.e., the distance between letters and signals in text, which typically comprises a large amount of ambiguity, redundancy, or irrelevant results. For example, when one searches for papers on the theory of SVM (support vector machine), most of the results from these systems may be papers about the application of SVM in other domains. In addition, the system based on bibliometrics usually analyses scientific papers from the document perspective. Specifically, it simply counts the absolute times that the terms occur, ignoring the deep semantic meaning and role of the terms in the paper, creating substantial noise in the process of literature analysis and yielding poor results.

In this article, we assert that taking the term function into consideration can solve the above problems in the existing systems. By recognizing the term function, we can utilize it to realize semantic retrieval, optimize the results of traditional academic databases, improve the experience of navigation and browsing function, and meet the user’s academic information needs much more precisely. In addition, by using term function recognition, we can
semantically organize and store scientific papers, which can provide a reliable base for further statistical work and analysis of scientific literature.

**Term function recognition (TFR)**

The definition of term function (TF) is the semantic role and specific function that a term plays in the scientific literature, given by Cheng (2015). In different academic texts, the same term may contain different semantic meanings. For instance, the term “SVM” plays different semantic roles in Chen and Fan (2009) and Zhang, Pan, Zhang, and Jiang (2004). In the former article, SVM is the research topic; whereas, in the latter article, it is the research method. Moreover, terms may have numerous functions, other than topic or method, such as goal, technology data set, etc., that are quite commonly used. Therefore, the classification and recognition of term function have constituted a popular and difficult task for researchers to solve because it is of critical significance for the automatic semantic analysis of scientific literature.

In the extant literature, there are few works directly relevant to term function recognition (TFR). The first study on TFR was conducted by Knodo, Nanba, Takezawa, and Okumura (2009). They divided terms in the title of scientific papers into “head”, “goal”, “method”, and “other” to discuss how to master the development trend of research technology from a specific research domain. Then, different scholars proposed their own classification of term function (TF), and conducted corresponding TFR experiments (see Table 1 for more details), but they did not give an explicit definition of term function. In Cheng (2015), term function was clearly defined and a more perfect framework of term function in academic text was proposed, which divided term function into “domain-independent term function” and “domain-related term function”. In his Ph.D. thesis, “domain-independent TF” has two types: “research topic” (including main topic, general topic, and other topic) and “research method” (including main method, general method, and other method). “Domain-related TF” mainly consisted of cases in economics, tools in computer science, data sets in data science, etc. (Cheng, 2015).

Overall, TF and TFR still constitute novel issues for scholars globally. Although several effective attempts have made, the topic remains largely undeveloped. In addition, an apparent gap exists between TFR and its application, which forms the exact initiation point of the present study.

<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Classification of term function</th>
<th>Recognition method</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>Kondo (2009); Nanba, Kondo, and Takezawa (2010)</td>
<td>Head, goal, method, other technology, effect</td>
<td>Sequence annotation Support vector machine</td>
</tr>
<tr>
<td>2011</td>
<td>Gupta and Manning (2012); Tsai, Kundu, and Roth (2013)</td>
<td>Focus, technique, domain Technique, application</td>
<td>Syntax templates Markov logic network</td>
</tr>
<tr>
<td>2013</td>
<td>Huang and Wan (2013); Cheng (2015)</td>
<td>Method, task, other domain-independent, domain-related Research topic Research method Case, tool, data set, etc.</td>
<td>Knowledge map Conditional random field Learning to rank</td>
</tr>
</tbody>
</table>

**Table 1. Classification and recognition of term function.**

**Design of literature analysis system**

Regarding the definition of term function (TF) proposed by Cheng (2015), we assumed that a scientific paper is a reflection of the process of several research topics solved by a number of research methods. On this basis, we recognize the research topic terms and research method
terms from each scientific paper to form a Topic Set, Method Set and Topic-Method Set, which can represent the main idea of the paper. It is important to note that, in this article, we only consider the main topic and main method of each academic work. In this way, we can semantically analyse scientific papers from the term perspective, which not only can provide researchers with a much more precise retrieval and navigation function, but can also more accurately detect and depict the trend and hot points of a scientific domain. In addition, by correlation analysis between the methods and topic extracted from the papers, we can realize a more intelligent and semantic scholarly recommendation system through method recommendation and topic recommendation.

The idea of literature analysis system

In order to implement a literature analysis system based on TFR, the first problem to be solved is to obtain a collection of scientific literature, and recognize and extract the main methods and main topic of each paper by utilizing the rule-based information extraction method. Secondly, we consider that the information needs of scholars are “to retrieve a research topic” or “to retrieve a research method”. To comprehensively fulfill these needs, we index, respectively, the methods and topics of each paper to provide a more fine-grained retrieval function for users, which can optimize the retrieval function of traditional academic databases. Regarding the retrieval results, the system can conduct further quantitative analysis on them, such as using the number of papers every year to depict the evolution of a topic or a method, and to detect the most commonly used methods and the hottest topic in a domain by the frequency that the methods or topic occur, etc. In addition, the system can also deliver a user-friendly interface and provide good visualization of the different results by using information visualization technologies. Finally, the system should be an interactive system, in which users can change the analysis dimensions and choose their own settings, such as time limitations, data updates, data management, etc.

Construction of literature analysis system

Considering the functions of similar existing systems and users’ needs, the system that we designed has six functional modules: 1) literature retrieval module; 2) navigation module; 3) research hot spot discovery module; 4) research trends analysis module; 5) correlation analysis module; and 6) data management module. The function framework of our literature analysis system is presented in Figure 1.

![Figure 1. Function framework of CS-LAS](image)

（1）Literature retrieval module

When searching scientific papers in traditional academic databases, it is very common for users to find results that are not what they actually wanted, a situation that is similar to the example proposed in section 2.2. In order to solve this major problem, the retrieval module of CS-LAS can index the main topic and main method extracted from each paper to provide users with a more precise retrieval function and optimize the results of traditional academic databases. The module comprises topic retrieval and method retrieval functions, which differ
according to the role of a user’s keywords in papers, and can be chosen by users with a checkbox next to the search box in the system.

(2) Navigation module
In cases in which a user may need to have a very wide scope of knowledge of a specific domain, the navigation module can be quite useful. The navigation module enables users to browse and analyse the topics discussed and methods used from the perspective of the entire domain perspective, including hot topics ranking, hot methods ranking, etc.

(3) Hot discovery module
Usually, scientists want to identify the hot topics and most common methods in a specific domain, the hottest and most reliable method for a specific topic, or the most common application areas for a specific method. The hot discovery module can fully satisfy these needs. On the one hand, the module can calculate the frequency of topics and methods, and rank them in a specific domain, and the results can show the hot rank list of different topics and methods in a specific domain. On the other hand, for a specific research topic, the frequency of corresponding method terms in the Method Set becomes the metric for hot discovery in this module, based on which we can obtain the most commonly used methods for a specific topic. In a similar manner, we can discern the hottest application area for a specific method.

(4) Trends analysis module
The main purpose of the trends analysis module in CS-LAS is to depict the evolution of methods for a specific topic or application trend of a specific research method. For a certain research topic, this module can calculate the document frequency of all methods for the topic (or all topics to which the method is applied), and obtain the time series sequences to present the trends.

(5) Correlation analysis module
In many cases, scientists would also like the system to recommend some methods that they may not be aware of that can solve the topic in which they are interested, or recommend new research topics to which methods that they are familiar with can be applied. This is why we designed the correlation analysis module for CS-LAS. By means of the collaborative filtering algorithm, this module can meet users’ needs. In the results of this module, detailed bibliographic information is also provided, and users can confirm whether or not the recommendations are appropriate for them.

(6) Data management module
The last module of CS-LAS is the data management module, which comprises three sub-modules: 1) system management; 2) spider management; and 3) extraction management. First, the system manager can modify and maintain the system through the web interface, such as modifying the path of the data files, data updates, etc. Second, the manager can set the period of data collection and the html parser by utilizing the spider management module. Finally, the extraction module allows the manager to add, delete, and modify the extraction rules.

Implications of literature analysis system
In this article, we take computer science as an example to realize a literature analysis system based on term function recognition (TFR), and name this as CS-LAS (i.e., Literature Analysis System for Computer Science). The implementation process of CS-LAS is shown in Figure 2, which can be divided into four steps: 1) the collection of raw data; 2) data processing, generating the Method Set, Topic Set and Topic-Method Set, and indexing them; 3) the implementation of retrieval and analysis modules; and 4) UI design and visualization of the results. This section will explicitly describe the issues that need to be addressed in each step.
Data collection
With the purpose of raw data collection for CS-LAS, we manually obtained the html pages of 307,455 articles published between 1994 to 2013 in the field of computer science from CNKI, which is the largest Chinese academic journal database. We then use an html parser, called Jsoup, and regular expression to parse the html pages and extract the bibliometric information of each paper, including the title in Chinese and English, publication year, author, abstract in Chinese and in English, keywords in Chinese and English, etc. Moreover, in order to adapt to the dynamic features of academic resources, the system can update data in an automatic and dynamic manner. In addition, the users can manually set the retrieval strategies and period of the data collection through the data management module, which can guarantee the accuracy and timeliness of the data set.

Construction of Topic Set and Method Set
For the construction of the Topic and Method Sets, we use extraction rules to extract functional topic terms or method terms from scientific papers that have been pre-processed. The extraction rules are generated by regular expression from the titles of the papers, which can be modified by users through the extraction module. The specific process of functional term recognition and extraction is as follows.

(1) Generation of extraction rules
Through observation of the Chinese and English titles of the academic papers, we found that there are a large number of titles written in the format of “基于 M 的 P” (an example is shown in Figure 3), which accounts for 28.9%. In the meantime, a title in this form can usually clearly express the main research topic and method of the paper, which inspired us concerning how to extract the main topic and method terms from the academic papers. In order to avoid the complexity of Chinese word segmentation and multi-expression of Chinese

Figure 2. Implementation process of CS-LAS.
terms, we decided to choose the English titles of papers, whose Chinese titles are in the format of “基于 M 的 P”, to be our experiment data set for the generation of extraction rules and functional term extraction.

After the pre-processing, including case conversion, word segmentation, de-symbolization and stemming, we represented the title of each paper as a word sequence $T = [W_1, W_2, W_3, \ldots, W_j]$. We also constructed a table of structural words $R = \{r_1, r_2, r_3, \ldots, r_j\}$ by observing the titles in the data set; every title can be treated as a word lattice, in which Chinese words like “基于” “的” “研究” and English word like “research” “on” “based” are structural words (an example shown in Figure 4). Next, we used a computer to automatically compare each word in $T$ with all words in $R$; if $W$ exists in $R$, then keep it and replace the others with “<word>”. Then, we use one “<word>” to replace all “<words>”, and can obtain the combination format of titles. Finally, we need to discern whether “<word>” in each format is a topic term or a method term manually: if “<word>” is a topic term, we replace it with “<Topi”c>”; otherwise, we replace it with “<Method>”. Table 2 shows the top eight combination formats of article titles in computer science and the corresponding extraction rules.

![Figure 3. An example for titles written in the format of “基于 M 的 P”.

![Figure 4. An example for word lattice of Chinese titles (a) and English titles(b)
Table 2. Combination format and corresponding extraction rules in computer science.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Combination format</th>
<th>Extraction rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt;word&gt; based on &lt;word&gt;</td>
<td>&lt;Topic&gt; based on &lt;Method&gt;</td>
</tr>
<tr>
<td>2</td>
<td>&lt;word&gt; based on &lt;word&gt; and &lt;word&gt;</td>
<td>&lt;Topic&gt; based on &lt;Method&gt; and &lt;Method&gt;</td>
</tr>
<tr>
<td>3</td>
<td>research on &lt;word&gt; based on &lt;word&gt;</td>
<td>research on &lt;Topic&gt; based on &lt;Method&gt;</td>
</tr>
<tr>
<td>4</td>
<td>a &lt;word&gt; based on &lt;word&gt;</td>
<td>a &lt;Topic&gt; based on &lt;Method&gt;</td>
</tr>
<tr>
<td>5</td>
<td>research of &lt;word&gt; based on &lt;word&gt;</td>
<td>research of &lt;Topic&gt; based on &lt;Method&gt;</td>
</tr>
<tr>
<td>6</td>
<td>&lt;word&gt; and &lt;word&gt; based on &lt;word&gt;</td>
<td>&lt;Topic&gt; and &lt;Topic&gt; based on &lt;Method&gt;</td>
</tr>
<tr>
<td>7</td>
<td>&lt;word&gt; of &lt;word&gt; based on &lt;word&gt;</td>
<td>[&lt;Topic&gt; of &lt;Topic&gt;] based on &lt;Method&gt;</td>
</tr>
</tbody>
</table>
| 8    | <word> based on <word> of <word> | <Topic> based on [<Method> of <Method>]

(2) Term function recognition and extraction

We used regular expression based on the extraction rules to match the topic and method terms with the English title of each paper. Finally, we obtained 30,145 Topic-Method relationships, which comprise 23,221 non-repeating topic terms and 18,427 non-repeating method terms. To verify the effectiveness of the extraction experiment, we randomly selected 1,500 results to test, and the accuracy is 97.6%. By indexing the papers with their topic and method terms, respectively, CS-LAS can provide users with much more precise retrieval and a better user experience.

Implementation of retrieval and analysis module

To implement the retrieval module, on the basis of indexing papers with topic terms and method terms extracted, we utilized classical information retrieval models, such as BM25, to retrieve papers and rank the results for users. The analysis component of this system contains three aspects: 1) the detection of trends; 2) hot spot discovery; and 3) correlation analysis. CS-LAS uses a statistical analyzer written in the Java programming language and classical data mining algorithms to analyze the Method Set, Topic Set, and Topic-Method Set to complete the three tasks.

User interface and visualization

CS-LAS employs a popular web-based visualization tool, called d3.js, to present the retrieval and analysis results for users. The main interface of the system comprises three parts: 1) retrieval interface; 2) navigation interface; and 3) analysis interface. Regarding the retrieval interface, when a user enters the query into the search box, he or she can press an interactive button to choose to retrieve the “research topic” or “research method”. Figure 5 shows the result of the retrieval of a research topic, i.e., “image segmentation”. In the upper-left corner of the picture, the most common methods used for solving this topic are presented, which are represented by a dynamic word cloud. Specifically, the greater the document frequency of a research method term, the deeper the color of the term in the word cloud. When the term is clicked on, the corresponding detailed bibliometric information about that article is shown. In the upper-right corner of the picture, the distribution of documents on this research topic is shown, i.e., the number of academic works about this research topic in each year. The bottom of this picture is a dynamic presentation of the papers in the upper-right corner, which contains the title, each paper’s year of publication, and a hyperlink to detailed bibliometric information.

In the navigation interface, users can browse all topics discussed and all methods used in the computer science article set. In the meantime, the system can automatically retrieve a research topic or a research method when it is clicked on.
Figures 6 and 7 are the visualization results of hot spot detection and correlation analysis. In Figure 6, users can control the analysis time limitation to generate the overview of hot topics and hot methods in a domain. The topics and methods are cross-represented on the structure of a hierarchy tree, in which the root of the tree represents that its son node is a topic (or a method), and the nodes in the next level indicate the methods used for (or the topics solved by) its parent node. When the node (except for the root node) is clicked on, the correlation analysis will be presented in the results, as shown in Figure 7. For example, when we click on the research method “SVM”, there will be three tables in the result pages. The first table is the topics in which SVM has been used, the second table is the methods in computer science that are similar to SVM, and the last table is the topics to which SVM may be applied. Each table has three columns: 1) rank, which represents the score obtained from recommendation algorithms; 2) the name of topics (or methods); and 3) the corresponding literature.

**Figure 5.** Results of the retrieval of a research topic: “image segmentation.”
Figure 6. Visualization of hot topics and hot methods in computer science from 1994-2003.

Figure 7. Correlation analysis results of research method “SVM”.
Discussion and Conclusion

In this paper, we presented CS-LAS, a novel literature analysis system that implements semantic analysis of scientific literature from the perspective of term function recognition (TFR). CS-LAS is a tool that may be quite useful to detect hot spots and to depict evolution trends of a specific academic domain, to recognize the most popular method for a specific topic, to recommend a possible method for an existing topic, etc. It can also provide much more precise retrieval and navigation for users. In the future, we plan to construct a larger Topic-Method Set, which might facilitate a major improvement of the analysis module. In addition, the recognition of term function in academic text that we conducted by a rule-based approach constitutes the key point of this study. To improve our system, an exploration of new methods, such as machine learning and deep learning, should be performed in future work.

Acknowledgments

The work described in this paper was fully supported by a grant from National Natural Science Foundation of China(No.71473183). We are indebted to our friend, Liu Xingbang. We are also grateful to the CNKI for their providing the experiment data set.

References


Spangler, S., Wilkins, A. D., Bachman, B. J., Nagarajan, M., Dayaram, T., & Haas, P. (2014). Automated hypothesis generation based on mining scientific literature. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp.1877-1886). ACM.


Incidental or influential? – A decade of using text-mining for citation function classification.

David Pride and Petr Knoth

The Knowledge Media Institute, The Open University, Milton Keynes, UK.
{david.pride, petr.knoth}@open.ac.uk

Abstract
This work looks in depth at several studies that have attempted to automate the process of citation importance classification based on the publications’ full text. We offer a comparison of their individual similarities, strengths and weaknesses. We analyse a range of features that have been previously used in this task. Our experimental results confirm that the number of in-text references are highly predictive of influence. Contrary to the work of Valenzuela et al. (2015) [1], we find abstract similarity one of the most predictive features. Overall, we show that many of the features previously described in literature have been either reported as not particularly predictive, cannot be reproduced based on their existing descriptions or should not be used due to their reliance on external changing evidence. Additionally, we find significant variance in the results provided by the PDF extraction tools used in the pre-processing stages of citation extraction. This has a direct and significant impact on the classification features that rely on this extraction process. Consequently, we discuss challenges and potential improvements in the classification pipeline, provide a critical review of the performance of individual features and address the importance of constructing a large-scale gold-standard reference dataset.

Conference Topic
[i] Citation and co-citation analysis [ii] The theory, method and principle of five metrics science concepts, that is, Bibliometrics, Informetrics, Scientometrics, Webometrics and Knowledgometrics. [iii] Methods and techniques.

Introduction
Citation analysis and bibliometrics are being increasingly used as a tool in assessing the impact of research. The three largest citation databases; Google Scholar, Web of Science (WoS) and Scopus all give prominence to citation counts to provide information regarding the number of times a paper has been cited. Most measures widely used to measure performance of research, such as the controversial Journal Impact Factor (JIF) [2], h-index [3] and Eigenfactor [4], rely on citation counts. All of the above methodologies suffer from the same base limitation in treating all citations equally.
It has been long established that treating all citations with equal weight is counterintuitive. Garfield, the original proponent of the JIF [2], proposed a range of 15 different reasons a paper may be cited [5]. These can include such reasons as: paying homage to pioneers, substantiating or refuting the earlier work of others, identifying methodologies used or simply giving background information regarding previous work. It can be seen from Garfield’s original list that simply counting citations cannot paint an entire picture of a papers impact.
Therefore, there is an increasing need in the automatic identification of the nature of a particular citation. Additionally, the growing availability of publication full texts is now making it possible to extend bibliometric studies further than those previously attempted with analysis of abstracts and citation networks alone. Open Access repositories such as that provided by
CORE1 [6] are allowing researchers to utilise the full text of research papers and articles in ways not possible with the meta-data offered by bibliographic databases alone. This has given rise to new areas of study including Semantometrics [7] which attest that the full text of a publication is required to effectively ascertain its impact.

In this paper, we address the problem of identifying influential citations based on publications’ full text. The rest of the paper is organised as follows. In Section 2, we introduce key studies on which our work is based. We then discuss the approach for detecting influential citation, providing a critical analysis of features previously applied in this task in Section 3, selecting a set of three key features for further analysis. We present a comparative study of the identified features in Section 4, together with the challenges inherent in this task.

Related Work

There have been several different methodologies applied to this task, Teufel (2006) [8] focuses on semantic similarity, identifying cue-phrases in the citing paper such as “we used” or “further to the work of”. Citations are classified into 11 types, which are then grouped into the higher grain categories of weak, positive and neutral sentiment. Later studies expand this and, rather than classifying only according to sentiment, attempt to classify citations as either influential or non-influential.

Hou et al. (2011) [9] first suggest the idea of using an internal citation count based on the full text of a research paper rather than just the bibliography to determine influence. They demonstrate a positive correlation between the number of times a citation occurs and its overall influence on the citing paper. Zhu et al. [10] combine these earlier approaches and suggest a range of 40 classification features including both semantic and metric features to determine influence. Most recently, [1] made significant efforts to construct a reference set which was publicly released and which this study relies heavily on. They suggest a range of 12 features, many of which show similarity with those of [10]. The features in these studies can be divided into having internal reliance or having external reliance. The former requires only the full-text of the citing paper whereas the latter relies on additional, external information being available. Furthermore, these studies identify three essential feature types. They are semantic based features, similarity based features and metrics / count based features. All of the studies under consideration use a range of different features and test them on different datasets. Consequently, getting a deeper understanding of which of the previously suggested features are most effective at this task is needed.

Methodology

- The typical workflow for classifying citation types involves the following steps:
  - Extracting the full text of the manuscript.
  - Parsing the full text to detect the document structure, such as the document metadata, references, citation markers and sections.
  - Extracting the features from the document structure, possibly with an enrichment step for features based on external evidence.
  - Applying a classifier trained using supervised machine learning approaches. In the rest of this section, we describe this workflow concentrating on the selection of features used in the citation type classification task.
Extracting the full-text and parsing

Unless a paper is available in a structure format, such as an XML, there is a requirement for converting the original PDF file into full text prior to analysis. There are numerous tools available for the conversion of PDF to text files. However, automatic text extraction from PDF is known to be problematic [11]. Some tools for inferring the document structure, such as ParsCit [12], require initial conversion to plain text. Others, such as GROBID [13], operate directly on the PDF file.

Features used by prior studies

One of the overriding aims of this work is to establish which of the previously identified classification features perform most strongly as predictors of citation importance and to use this as a baseline from which to build future work. We consider the features presented in the two most recent studies. In [10] we first see an expansion of the features into a rich range that move beyond simple counting of in-text citations;

- Count-based features
- Similarity-based features
- Context-based features
- Position-based features
- Miscellaneous features

Valenzuela et al. (2015) [1] take a similar approach to the construction of the features list. The 12 features used in this study are;

- F1 Total number of direct citations
- F2 Number of direct citations per section
- F3 Total number of indirect citations and number of indirect citations per section
- F4 Author overlap (Boolean)
- F5 Citation Is considered helpful (Boolean)
- F6 Citation appears in table or caption
- F7 1 / Number of references
- F8 Number of paper citations / all citations
- F9 Similarity between abstracts
- F10 PageRank
- F11 Number of citing papers after transitive closure
- F12 Field of cited paper.

Selection of features for experiments and comparison

We analysed the 40 features presented by Zhu et al. [10] and 12 features presented in the study of Valenzuela et al. [1] Of the 40 features, a combination of just 4 features resulted in the best performance of Zhu's model. Adding features beyond this actually lowered the performance. Out of these 4 features, we could not reliably replicate one feature (countsInPaperSecNum). Out of the 12 features of Valenzuela (Table 1) we found three features irreproducible (F3, F5 - We attempted to reproduce this feature, but failed due to Valenzuela's dictionary of cue words not being available, F12), we were unable to reliably replicate two features due to PDF extraction issues (F2, F6) and we elected not to use two features as they rely on external and
potentially changing evidence (F10, F11). Two features we tested (F7, F8) did not produce any significant correlation with the gold standard.

Of the three remaining features of Valenzuela, we found a complete overlap of two features (F1-countsInPaperWhole, F4-aux\_SelfCite) and a close match on the third (F9-simTitleCore). These three selected features correspond to the best (F1-countsInPaperWhole) feature of Zhu, the worst feature of Valenzuela (F9-simTitleCore) and a third where the opinion regarding the usefulness of this feature was divided between the two studies (F4-aux\_SelfCite). In the rest of the paper, we will provide a direct cross-comparison of these features on a single dataset:

**Number of direct citations (Integer):** This feature is labelled by [1] as 'F1 - Direct Citations' and by [12] as 'countsInPaperwhole.' Both of these studies, and the earlier study by [hou2011counting] find the total number of times a paper is cited to be a strong indicator of academic influence on the citing paper.

**Abstract Similarity (Real):** This is feature F9 in the study by [1]. Whilst [10] tested various similarity based features, none performed better than their randomly assigned baseline (equivalent to the prior distribution of the influential label in their dataset). Valenzuela et al. [1] also listed this as the weakest feature. This feature is calculated as the tf-idf cosine similarity between citing paper abstract and cited paper abstract.

**Author Overlap / Self-Citation (Boolean):** This feature is labelled F4 by [1] and as [10]. The two studies differ markedly in their opinion of the value of this feature. While [11] found little correlation between author overlap and influence, [1] listed author overlap as their third best performing feature.

**Classification**

Using the identified features, we perform a binary incidental / influential classification. WEKA 3 [14] was selected as the machine learning toolset in our study.

**Results**

**Dataset**

The dataset released by [1] contains incidental/influential human judgments on 465 citing-cited paper pairs for articles drawn from the 2013 ACL anthology, the full texts of which are publicly available. The judgment for each citation was determined by two expert human annotators and each citation was assigned a label. Both a fine-grained (4-Way) label and a binary (incidental / important) label were provided. Using the author’s binary classification, 396 citation pairs were ranked as incidental citations and 69 (14.3\%) were ranked as influential (important) citations.

It is extremely interesting to note that all studies which employed human annotators to judge citation influence [1,8,10,15] reported a broadly similar ratio of positive examples. This ranged from 10.3\% [10] through 14.3\% [1] to 17.9\% [8]. This is an important finding as it gives a clear indication that only a relatively small percentage of all citations are actually influential at all. All of the studies find that the majority of citations are perfunctory at best. Negative citations are extremely rare and this in itself further increases the difficulties in constructing a balanced reference set. Automatic identification of those influential citations is therefore both a more important and less straightforward task than may be first imagined.
To obtain a clean dataset for our experiments, we first collected the PDF files of the citing and cited papers used by Valenzuela et al. [1] from the ACL Anthology. We processed these papers using pdf2txtt [16] to extract metadata, citations, the full text and other document structure information. Any papers where the extraction was not possible or the abstract was not available were then removed. This left us with a dataset of 415 pairs with 355 citation pairs marked as incidental and 60 citation pairs (14.45%) marked as influential citations. As this corresponds to only a relatively small reduction in the number of examples from the original dataset and reflects the original ratio between incidental and influential citation classes, we consider this dataset to be sufficiently representative for our experiments. We then process the XML files using ParsCit and applied calculations to extract features for each example.

**Analysis and comparison of selected features.**

Our experiments tested a range of features and their efficacy as predictors of citation influence. We achieved the best results using the Random Forests Classifier. We tested the model using bagging with 100 iterations and a base learner, using a 10-fold cross-validation methodology. The WEKA toolset was used to generate P/R curves for each of the individual features as well as the combination of all the features (Table 1).

<table>
<thead>
<tr>
<th>Feature</th>
<th>P@R=0.05</th>
<th>P@R=0.1</th>
<th>P@R=0.3</th>
<th>P@R=0.5</th>
<th>P@R=0.7</th>
<th>P@R=0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>0.4</td>
<td>0.34</td>
<td>0.33</td>
<td>0.3</td>
<td>0.26</td>
<td>0.21</td>
</tr>
<tr>
<td>F4</td>
<td>0.27</td>
<td>0.35</td>
<td>0.14</td>
<td>0.15</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>F9</td>
<td>0.46</td>
<td>0.49</td>
<td>0.21</td>
<td>0.2</td>
<td>0.18</td>
<td>0.16</td>
</tr>
<tr>
<td>All</td>
<td>0.5</td>
<td>0.38</td>
<td>0.37</td>
<td>0.37</td>
<td>0.29</td>
<td>0.23</td>
</tr>
</tbody>
</table>

We also measured the correlation between each of the individual features and the classification given by the human annotators. Valenzuela et al. [1] present their results in terms of P/R values for each feature whereas [10] shows the Pearson correlation with their gold standard. We therefore present the results of our experiments in both formats to allow for accurate comparison. Our work confirms the earlier findings reported in [10] and [1] that the number of direct instances of a citation within a paper is a clear indicator of citation influence. We also find that author overlap, or self-citation, does have value as a classification feature. Contrary to the work of [1] we find that the similarity between abstracts is more predictive of citation influence than previously shown.

**Table 2. Comparison of results by feature**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Valenzuela et al.</th>
<th>Our results</th>
<th>Zhu et al.</th>
<th>Our results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Citations</td>
<td>0.3</td>
<td>0.21</td>
<td>0.33</td>
<td>0.281</td>
</tr>
<tr>
<td>Abstract Sim.</td>
<td>0.14</td>
<td>0.14</td>
<td>N/A</td>
<td>0.373</td>
</tr>
<tr>
<td>Author Overlap</td>
<td>0.22</td>
<td>0.16</td>
<td>0.02</td>
<td>0.132</td>
</tr>
</tbody>
</table>

The correlation of this feature with the reference set (r=0.373, p < 0.01, 2-tailed) was the highest of all the features we tested. It is our contention that testing all features using P/R
values, at R0.90 masks some of the predictive value of those features when the dataset contains only a small number of instances of the influential class. Table 2 shows the precision of the random forests classifier at various recall levels. It can be seen from these results that the classifier initially performs quite well and identifies many of the influential cases, however it has difficulty identifying the last few instances which substantially decreases the classifier’s performance at R0.90. Using Mean Average Precision (MAP) or a similar metric that provides a single-figure measure of quality across recall levels would be a better choice in this case.

Results for Individual Features

Number of Direct Citations - F1: This feature is rated as the highest value in terms of predictive ability by [10] and the second highest by [1]. The latter shows P0.30 at R0.90, however our results demonstrate a slightly lower P value, P0.21 at R0.90.

Zhu et al. [10] list the equivalent ‘countsinPaper_Whole’ as the most significant feature of their classifier, with a Pearson correlation coefficient of P0.35. We find a Pearson correlation of P0.28 (significant at the 0.01 level, 2-tailed) for this feature with our dataset. The small difference in this result is likely caused by the differences in the two datasets. Our results therefore confirm that the number of times a citation appears is a strong indicator of that citation’s influence.

Author Overlap - F4: The results from the two earlier studies for this feature vary considerably. In the results for [1] this is the third ranked ‘most significant feature with P0.22 for R0.90. We find slightly less precision than [1] for this feature; P0.16 at R0.90. [10] results show little correlation with their gold standard for the similar feature aux_selfCite (Pearson 0.02). Interestingly, despite the low correlation, this feature was the fourth one selected by their model and
did indeed improve the performance of the classifier, albeit only slightly. The experiments with our dataset show a far stronger positive correlation, P0.132 (significant at the 0.01 level, 2-tailed), than that found by [10].

Abstract Similarity - F9: Whilst [10] generated many similarity-based features, they did not compare citing abstract and cited abstract. This is somewhat surprising as we consider it to be an interesting feature and one that also seems innately logical. The abstract similarity is calculated as the cosine similarity of the tf-idf scores of the two abstracts. By ensuring that the dataset only contains valid data, i.e. the abstract is available for both citing and cited paper, a direct comparison can be made for this feature with [1] who rank this as the lowest of their twelve features, P0.14 at R0.90.

Here our results are the same as [1], with P0.14 at R0.90. However, the Pearson correlation with the gold standard dataset for this feature is the highest of the three features tested in our experiments. We find a Pearson correlation of 0.373 (significant at the 0.01 level, 2-tailed). This feature was not tested by any of the other earlier studies covered in this work. Our results demonstrate that abstract similarity between citing and cited paper is more predictive of citation influence that previously shown.

The value of complex features.

It is now over a decade since Teufel [8] first attempted to automate the classification of citation function. This original study and several subsequent ones have suggested classification features ranging from the extremely simple to the extremely complex. Many of these complex features have been shown to have little predictive ability in regards to classifying citation function or importance. Some of the most basic features have been shown to offer the strongest potential
in identifying important or influential citations. Our research confirms that one of the most simplistic features, i.e. the number of times a citation appears in a paper, is highly predictive of influence. Replicating complex features is a non-trivial task unless exact details of how the values for these features were calculated or source code are provided by the original study. We believe that it is essential that the types and values of all features should be provided as part of the research dataset (as opposed to providing just source prior to feature extraction) to serve as a roadmap in replicating them. Furthermore, features that rely on external datasets, changing evidence (e.g. citations, downloads, etc.), or utilise sets of rules, that are not available to other researchers cannot be replicated by this or any other future study. There has now been a decade of research in this area and the predictive ability of many complex features is still uncertain. This is sorely detrimental to the overall value of the original studies.

Analysis of PDF extraction.

Both [1] and [10] use ParsCit - the citation parsing tool, based on Conditional Random Fields (CRF). As this is a critical pre-processing stage we conducted experiments to determine the efficacy and accuracy of this tool. GrobID [13] is a similar CRF based tool and was chosen to provide a comparison.

There are several types of errors that can be introduced during the PDF conversion process:

- PDF is a scan and would require OCR.
- Custom encoding instead of Unicode or ASCII.
- Readable XML file not created at all due to failed PDF conversion process.
- References not identified or counted correctly.
- Citations not identified or counted correctly.
- Abstract not extracted correctly or not present in cited paper.
- Title names / Author names misspelled in different parts of the paper.
- Elements being mis-tagged.

We argue that these errors unavoidably impact on the validity of any classification features that are reliant on this process. Of particular concern is the likelihood of citations being either under-counted or over-counted. The results of our experiments demonstrate that this is indeed the case in many instances. To understand the impact of this, we conducted the following experiment. Ten papers were randomly chosen from the Valenzuela dataset and the citation counts for each citation were extracted using both ParsCit and GrobID. The results of both tools were then compared to a manual check / count. [1] and [10] demonstrate that the number of times a citation appears in the body of the text is a significant indicator of influence. There is however a difference in the number of citations identified, depending on the chosen method of parsing. The reference count for five of the chosen example papers is shown in table 3.

These results show that ParsCit correctly identified the exact number of citations in only 40% of cases. GrobID was even less successful. It was exactly correct in only one case and missed a significant number of citations in many others. We argue that this demonstrates a potentially serious failing in current methodologies that rely on PDF extraction for calculation of number of citations.
Table 3. Comparison of in-text citations counts by extraction method

<table>
<thead>
<tr>
<th>Paper ID</th>
<th>ParsCit count</th>
<th>GrobID count</th>
<th>Actual Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>C00-2140</td>
<td>33</td>
<td>21</td>
<td>35</td>
</tr>
<tr>
<td>W06-0202</td>
<td>17</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>W09-1118</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>E12-1072</td>
<td>31</td>
<td>21</td>
<td>30</td>
</tr>
<tr>
<td>P02-1058</td>
<td>13</td>
<td>8</td>
<td>13</td>
</tr>
</tbody>
</table>

Discussion

One of the major limitations of this and previous studies is the size of the publicly available, annotated, datasets. The study by [1] uses 465 citing / cited paper pairs. The study by [10] uses just 100 papers by 40 authors. Due to the unbalanced split between the incidental and influential classes, our complete dataset contained only 61 examples of the positive (influential) class. We argue that due to the relative sparsity of influential citations a much larger reference set is required. This is equally true for negative citations, which have been shown to be even rarer. Training a classifier when the dataset contains so few instances of the non-neutral classes is problematic and we will address this in future work. The construction of a gold standard dataset containing many thousands of annotated citations, rather than a few hundred, is a significant undertaking but we believe this is a vital step in improving the abilities of the classification models.

There is a noticeable difference between the datasets used by [10] and [1] which warrants further study. The [1] dataset annotation was undertaken by two independent annotators and finds significant value in using author overlap as a classification feature. However, the [10] reference set is annotated by the authors themselves and this study ranks author overlap / self-citation as being of very low importance. It may be that is demonstrates shyness or reticence on behalf of authors to regard their own, earlier, work as being a significant influence. A large-scale author-annotated reference set would be extremely helpful in ascertaining the level of this bias when compared to an anonymously-annotated dataset such as that of [8] or [1]. Finally we argue that if a citation is considered influential, this original influence remains regardless of external factors or the environment. Therefore, classification features which rely on external and potentially fluid information should be used somewhat cautiously. In future work we will address this issue in greater detail.

Conclusions

Of the features we tested we find Abstract Similarity shows the strongest positive correlation for predicting citation influence. We find Number of Direct Citations to also be highly predictive and we find Author Overlap / Self-Citation to be less predictive but still valuable as a classification feature. It is important to note that many of the features suggested by earlier studies have been shown to have little predictive ability. Additionally, despite significant efforts, we were not able to reproduce or validate several of the features used by [1] or by [10].

Furthermore, our results demonstrate that any automatic classification model that relies on PDF extraction in the pre-processing stage is unlikely to capture all of the relevant data which is fundamental to the calculation of the value of some features. We argue that this introduces a level of potential inaccuracy that has not been fully addressed. There is scope for further work surrounding the efficacy and in particular the reproducibility of some of the previously tested
classification features. Many of the earlier studies in this domain present results based on sometimes complex and irreproducible features. We contest that this is detrimental to this area of study as a whole and, whilst earlier studies have identified several effective features, having the ability to reproduce them is fundamental to further development in the area of citation classification.

Whilst it may be a relatively easy task for a human being to identify important or influential citations, building a model to automatically classify these citations with any degree of accuracy is a non-trivial task. A larger scale reference set than those used in this and previous studies is essential, particularly due to the inevitably skewed nature of any dataset of citations annotated according to influence or importance.

Acknowledgements

This work has been funded by Jisc and has also received support from the scholarly communications use case of the EU OpenMinTeD project under the H2020-EINFRA-2014-2 call, Project ID: 654021

References


Garfield, E., et al.: Citation analysis as a tool in journal evaluation, American Association for the Advancement of Science (1972)

Hirsch, J.E.: An index to quantify an individual’s scientific research output. Proceedings of the National academy of Sciences of the United States of America (2005) 16569–16572

Bergstrom, C.: Scholarly communication. Eigenfactor: Measuring the value and prestige of scholarly journals


A Comparative Investigation of Citation Content Characteristics between Academic Monographs and Papers

Wang Yuzhuo1, Zhang Chengzhi2,*

1 njustwyz@163.com
Nanjing University of Science and Technology, Nanjing (China)

2 zhangcz@njust.edu.cn
Nanjing University of Science and Technology, Nanjing (China)
Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University), Fuzhou (China)

Abstract: Cited frequency of academic papers (APs) or academic monographs (AMs) are used to evaluate their academic influence usually. This frequency-based method does not consider detailed citation information, e.g., citation content or context, in citing paper. Citation context or content can reveal different citing motivation and behavior of authors. Citation content analysis (CCA) of APs or AMs can improve performance of their influence assessment. AMs are different from APs and they have characteristics in many ways. If we know the difference of citation content between AMs and APs, a more appropriate CCA method for AMS will be explored based on the existing CCA method of APs. Therefore, a comparison of citation content between AMs and APs is investigated in this paper. We annotated 39 academic monographs with 13, 539 citation sentences and 7, 756 academic papers with 590, 740 citation sentences respectively. Furthermore, we compared the location, mention times and length of citation content between AMs and APs and analyzed cause of the difference. The results quantitatively show that there are differences in location and length of citation between AMs and APs, while citation mention of the two is similar.

Keywords: characteristic of citation content, academic monograph, academic paper

Conference Topic
Knowledge discovery and data mining

Introduction
Citation analysis includes citation bibliographic analysis and citation content analysis. Traditionally, citation analysis mainly used the bibliographic information including references of academic papers to assess academic influence of cited papers. Currently, more and more full-text of academic papers can be accessed via academic database, and it’s easy for us to obtain citation content of references in a paper. Many researchers began to analysis citation content and use it to academic influence assessment (Hu, Chen & Liu, 2013; Zhang, Ding & Milojevic, 2013; MasakiEto, 2013; Wan & Liu, 2014; Bertin, Atanassova, Larivière & Gingras, 2015; Tahamtan et al,2016). At present, citation content analysis (CCA) is mainly based on academic paper according to three aspects: syntactic, semantic and pragmatic (Small, 1982; Chang, 2013; Jeong, Song & Ding, 2014). To the best of our knowledge, this paper is the first work that analyze citation content of academic monographs. Compared with journals, updating of academic monographs has a time delay, and monographs have richer and longer contents. Besides, the lack of full text data makes it difficult to extract large-scale citations from academic monographs. So, there is no relevant research on monographs’ citation content. At the same time, APs and AMs are both important knowledge carrier. What is the difference or similarity between their citation characteristics? What is the reason for these difference or similarity? These questions haven’t been investigated in the existing

* Corresponding author, Email: zhangcz@njust.edu.cn
works. Based on this study on the difference of citation content between AMs and Aps, we would explore a fit CCA method for AMs. Therefore, a comparison of citation content between AMs and APs is investigated in this paper. 39 academic monographs (APs) and 7,756 academic papers (APs) are chose. We compare their citation content from three aspects: location, mention times and length, and find AMs has similar citation mention to APs, but location and length of citation of the two are different.

Related works

Citation content analysis of academic papers

Citation content analysis of academic papers contains three levels: syntactic, semantic and pragmatic. From the syntactic point of view, scholars mainly study the citation location, citation mention and length of citation by analysing the full text of academic papers. As early as 1976, Voos and Dagae (1976) conducted a research of the citation location. They selected four highly cited papers and found that citations at different positions had different value and purpose. Highly cited papers are more prone to appear at “Introduction” section. On that basis, Ding et al. (2013) adopted a new method, namely “count X”, and found that highly cited papers were more gathered in “Methodology” section. In general, the first 10%~15% of the full text was the densest filed where citations located, and half of citations were located in the first 30% of the full text (Halevi & Henk, 2013). In four-section (IMRC) papers, citations were more likely to locate in the section of Introduction. While inside section, citations were distributed evenly and randomly (Hu et al, 2013). Studying the position can help scholars to understand the distribution of citations in the full text. Statistics of the cited frequency of a reference can be used as the measure of its influence. By traditional methods only the number of times which the paper appears in the reference list were counted. Zhu (2013) built four different influence prediction models and observed that the times of a paper mentioned in the full text (count X) was the best standard to measure impact of the paper. Regarding to the average number of citations mentioned in a paper, the results of the early studies are low——for about 1.10 and 1.20 (Bluma C, 1983; Oppenheim & Renn, 1978). In recent years, mentioned times of general papers have risen to 1.50~1.79, and that of highly cited paper is 1.90~2.00 (Ding et al, 2013; Wan & Liu, 2014). There are few researches refer to the length and the number of words of citation sentences. A small amount of results show that the average number of words of citation is 26, and the length is mainly in the range of 0~200 characters.

In addition to the syntactic analysis, Scholars also carried out semantic research on citation content. As early as 1986, Small (1986) used the topic word in citation content to reveal the co-citation network clustering theme. Nanba and Okumura (2005) collected all the citation content of a paper, and described the main content of the paper by analyzing keywords of the citation. The researchers mainly studied the difference between the subject form by citation content and the abstract of the citation itself, and did not analyze the reason of the difference.

Pragmatic analysis mainly explored the emotion (Small, 2011), functions (Lin et al, 2014) and motivation of citations. Garfield (1964) first proposed 15 motivations of citation. Small (1982) and White (2004), respectively, explored the motivation in the field of information science and linguistics. Weinstock (1971) used the citation content to systematically summarize the motivation of citation and pointed out 15 motivations...
for scientific communication. However, these work only categorized the motivation and do not investigated its value in citation analysis.

Citation analysis of academic monographs

Scholars usually have to cite plenty of references during the writing of their monographs. Researchers carried out a series researches on the citations of monographs. Statistics of the source of citations in books indicated a complementary relationship between the three academic databases, i.e. WOS, Scopus, Google Scholar (Bar-Ilan, 2010). Torres-Salinas (2013) mapped citation patterns of book chapters in the BKCI and observed that the citation patterns of book chapters followed a Lotkaian distribution. Hammarfelt (2013) mined the interactive mode of disciplines and knowledge structure in humanities and social science filed based on the analysis of highly cited monographs. Some scholars compared academic monographs with academic papers at the same time to measure the value of academic monographs in academic exchanges. Thompson (2002) investigated the source and destination of citation in academic monographs and found that academic monographs played an important role in examining publishing models. Kousha and Thelwall (2009) analysed the same papers that were cited in different books and papers, the results indicated the book in Google Book Search could make up for the book citation shortage in ISI. On this basis, Kousha and Thelwall (2013) used the new Google Book citation automatic extraction method and compared the extractive results with BKCI, which showed there was a very high correlation between the two, even Google Book performed better in some areas. Chi (2016) analyzed the data of BKCI and observed that the referenced influence of humanities books and that of journals were quite different, while the influence of social science books and that of journals were similar.

In a word, existing works of citation content analysis at syntax level still mainly target academic papers. By using bibliographic information in academic monographs, most of citation analyses of academic monographs only focus on data source, transmission pattern and so on. There is not any woks that investigate citation content characteristics of monograph.

Methodology

In this paper, comparative analysis of citation content (CC) between AMs and ACs according to location, mention times and length of citation content respectively.

Dataset

The dataset contains 39 academic books1 from Morgan & Claypool2 and 7,756 papers from Public Library of Science (PLOS). The academic monographs are in PDF format and academic papers are available in XML format, both of them are related to two disciplines——Biology and Computer Science.

Data Annotation

We extract citation content of 39 monographs manually, and 7,756 articles automatically respectively. Meanwhile, other relevant information is annotated, such as citation context, references, and chapter numbers. In this paper, citation context includes the first and second sentence previous to citation sentence and the first and second sentence after

---

1 All of the 39 books were manually judged as academic monographs. A detailed explanation of the data scale is given in Section 5.1
2 http://www.morganclaypool.com/
citation sentence. Section means the chapter in AMs and the section in APs. A citation
refers to a sentence that contains quote marks, and the reference means the cited article
in the citation. One citation could contain more than one reference. Meaning of labels
is shown as Table 1. We obtain 13, 539 and 590, 740 citation sentences from AMs and
APs respectively.

Table 1. Meaning of Labels in the annotation corpora

<table>
<thead>
<tr>
<th>Name</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>Unique identifier of a monograph or an article</td>
</tr>
<tr>
<td>Citation_id</td>
<td>Sequence number of each citation sentence</td>
</tr>
<tr>
<td>Section_id</td>
<td>Sequence number of each section</td>
</tr>
<tr>
<td>Word_total</td>
<td>Total number of words of a monograph or an article</td>
</tr>
<tr>
<td>Word_location</td>
<td>Position of first citation’s word in the full text</td>
</tr>
<tr>
<td>Section_total</td>
<td>Total number of words of a section</td>
</tr>
<tr>
<td>Section_location</td>
<td>Position of first citation’s word in the section</td>
</tr>
<tr>
<td>Citation_content</td>
<td>Content of the citation sentence</td>
</tr>
<tr>
<td>Previous_1</td>
<td>The first sentence previous to citation sentence</td>
</tr>
<tr>
<td>Previous_2</td>
<td>The second sentence previous to citation sentence</td>
</tr>
<tr>
<td>After_1</td>
<td>The first sentence after citation sentence</td>
</tr>
<tr>
<td>After_2</td>
<td>The second sentence after citation sentence</td>
</tr>
<tr>
<td>Reference</td>
<td>The reference (cited article) of this citation</td>
</tr>
<tr>
<td>Reference_total</td>
<td>The total number of reference in the reference list of a monographs or an article</td>
</tr>
</tbody>
</table>

Analysis of three aspects of CC
Using citation content and its related information from previous steps, we make statistics
about the characteristics of CC in AMs and ACs respectively.
(1) Location of CC
We get relative positions of each citation in the full text by "word_location / word_total",
and calculate "section_location / section_total" to obtain relative position of each
citation in each section.
(2) Mention of CC
Just to be clear, if an article contains 10 citation sentences and each citation contains 3
references, and there are 20 references in the reference list of this article. Then number
of citations of this article is 30, and number of reference of this article is 20.
First of all, in terms of full text, we count number of citations and references of all the
papers and monographs based on the “citation_id” and “reference_total”, and calculate
ratio of the two. After that, number of citations in each monograph and paper are counted
to calculate the ration of citation to reference. We call the ration “mention of citation
content”
In terms of chapter, number of citations are counted by each chapter, we remove
repeated reference to get the number of references and obtain the “mention of citation
content”. On this basis, we normalize each chapter and compare the mention of citation
in different chapters.
(3) Length of CC
In this article, we regard number of words as length of a sentence. In the end, we count
number of words in each citation sentence and its context respectively, and obtain the
average length of citation content and citation context. The average length of CC is taken
as standard to normalize all the average length of citation context. To discover the
difference in length of CC between monographs and papers, the Kolmogorov-Smirnov (KS) test is carried out showing that the results are highly significant (p<0.01) for all variables.

For the analysis of three characteristics in the ahead section, we organize them with tables and graphs, and compare the location, mention and length of citation content between AMs and AC.

**Comparison Results of Citation Content between AMs and APS**

*Location of citations*

*Location of citation content in full-text*

Distribution of citations in full text of 39 academic monographs is shown as figure 1. According to figure 1, we can see: distribution of citation contents in full text of academic monographs is uniform and there is no obvious concentrated field.

![Figure 1. Distribution of citation content of academic monographs](image1)

The relative location of each citation sentence of academic paper is shown as figure 2. According to figure 2, it is easy to find that citation contents are mainly distribute at both ends of papers while citation content in the middle of papers are sparser. In particular, the first 15% of full text is the densest field where citations locate.

![Figure 2. Distribution of citation content of academic papers](image2)

*Location of citation content in chapter and section*

Figure 3 displays distribution of citation content in chapter of academic monographs, where each column represents a chapter. In 39 monographs, citation content evenly
distributes among the chapter. There are no significant differences between different chapters, either.

Each column in Figure 4 represents each section of each academic paper. Distribution of citation content in section of academic papers is also uniform, but density of citations at the end of each section (after the relative position of 0.8 in the figure) is significantly sparser than the previous part.

Figure 3. Distribution of citation content in chapter of academic monographs
Figure 4. Distribution of citation content in section of academic papers

In conclusion, citation content evenly distributes in full text of monographs, but there is an obvious concentration of citation content in academic papers. Furthermore, citation sentences are evenly distributed in the different sections of both academic monographs and academic papers.

**Mention of citation content**

**Mention of citation content in full text**

Regarding 39 monographs as a whole, number of references is 8, 885 and number of citations in full text is 13, 539, mention of citation is 1.52, that is, each reference is averagely cited 1.52 times.

As for each monograph, amount of citation content, amount of references and ratio of the two is shown in Figure 5 and Figure 6 respectively. With increase of amount of references, there are many twists and turns in line of citation content. Therefore, amount of citations is not in direct proportion to number of references. But mention of citation content is more than 1 and it concentrates at 1 to 1.5 in most of academic monographs, which means at least there is a repeated citation in each academic monograph.

Figure 5. Amount of citations and references in monographs
Figure 6. Mention times of citation content of monographs

We obtain 370, 508 references and 590, 740 citations from 7, 756 academic papers, and the mention of citation content is 1.59. Figure 7 indicates that as amount of references increases, amount of citations is increase, but there are some fluctuations on the line, which also proves that the total amount of citations is not directly related to amount of references. And then we calculated ratio of the two, in 7, 756 papers, and only 28 papers’ ratio is 1, which shows that most of academic papers contain repeated citation. The ration of more than half of the papers concentrates in 1 to 1.5, and only 69 papers’ ration goes over 3, accounting for less than 1%. The overall distribution is shown in Figure 8.
Mention of citation content in section

Each spot in Figure 9 represents mention of citation content in each chapter of each academic monograph. We order them by relative chapter. It reflects that average mention of citation content in each chapter is mainly in 1.0 to 1.5. Most chapters whose mean mention of citation content beyond 2.5 are after the 0.5 chapter, which indirectly implies that highly cited references are mainly in the latter half part chapters of academic monograph.

In Table 2, the average mention of citation content is calculated per 10% of total chapters. Frequency of the first and the last section is obviously lower than others sections. And mention of citation content in different 10% chapter is irregular.

Each column in Figure 10 is frequency of citation mention times of each section of each academic paper. The most of frequency is still between 1 and 1.5.

Subsequently, we rank previous results by each 10% of total sections. The result is shown in Figure 11. If ration of citations to references reached 3, we think there are more highly cited references. So, in academic papers, a high time of citation mention is more likely to appear at the 40%-80% of total sections, that is the latter part of the paper.

Calculating average mention of citation per 10% of total sections, we could see average mention of citation in middle section is higher than the beginning and the end of the paper (see Table 3).

![Figure 7. Amount of citation content and references of academic papers](image1)

![Figure 8. Mention of citation content of academic papers](image2)

![Figure 9. Mention of citation content in relative chapter of monographs](image3)

<table>
<thead>
<tr>
<th>Relative chapter</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>mention/reference</td>
<td>1.08</td>
<td>1.32</td>
<td>1.37</td>
<td>1.38</td>
<td>1.38</td>
<td>1.49</td>
<td>1.37</td>
<td>1.46</td>
<td>1.30</td>
<td>1.20</td>
</tr>
</tbody>
</table>

Table 2. Average mention of citation content in relative chapter of academic monographs
In general, the mention of citation content of monographs is almost similar to that of papers. But unlike the irregularity of academic monographs, the median section of paper has a higher mention of citation when compared with the beginning and ending sections.

**Length of citation content and citation context**

<table>
<thead>
<tr>
<th>Length</th>
<th>citation</th>
<th>Previous_2</th>
<th>Previous_1</th>
<th>After_1</th>
<th>After_2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AP</td>
<td>29.22**</td>
<td>1.00</td>
<td>25.17</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>AM</td>
<td>26.39**</td>
<td>1.00</td>
<td>25.28</td>
<td>0.96</td>
</tr>
</tbody>
</table>

**Significant at p<0.01**

Original average length and normalized length of citation content and citation context are presented in Table 4. It is obvious that length of citation content between academic monographs and papers has significant difference (Z=-1.26E-301, p=0). Citation content in academic papers is longer than citation content of academic monographs, but average length of citation context of the two is similar.

As for citation content and citation context, average length of the sentences closes to citation content (previous_1, after_1) is slightly larger than that of the sentences far from citation content (previous_2, after_2), and the difference of length between citation content and citation context of academic papers is particularly more evident than that of the monographs.

**Discussion**

*The data scale in this investigation*

In order to investigate whether magnitude of samples have an effect on the results, we randomly select 8 academic monographs, 40 academic papers and 400 academic papers to analyze location, mention times and length with the same previous methods respectively.

Figure 12 and figure 13 indicate that location of citation in full text of 8 academic monographs is similar to 39 academic monographs, and location of citation in full text 40, 400, and 7,756 academic papers are same, too. What’s more, the sampling results of
mention times and length don’t have difference with original results (Detailed comparison is not showed due to space limitations of this paper). Therefore, the differences between data scale of AMs and APs won’t be an interference factor on the analysis results.

![Figure 12. Location of citation in full text between different scales of academic monographs](image)

**Figure 12. Location of citation in full text between different scales of academic monographs**

![Figure 13. Location of citation content between different scales of academic papers](image)

**Figure 13. Location of citation content between different scales of academic papers**

**Reason of the differences between AMs and APs**

From previous analysis, it is obvious that there are great differences between AMs and APs in location of citation content, which is closely related to the content arrangements of academic monographs and academic papers.

In general, the opening part of academic paper is introduction and literature review and contains a large number of citations. The middle of paper is a self-statement part of author which don’t need too many citations. Some authors may compare their own work with others’ research results and even move the related work to the end of paper. As for academic monographs, contents of whole book are interlinked but the authors may treat each chapter as an independent unit and cite reference evenly. So, the overall distribution of citations in academic monographs presents a more homogeneous state, however, there is an obvious concentration of citation in academic papers.

**Suggestions to citation content analysis of academic monographs**

The same distribution of citation content in different sections suggests that we have no need to consider position of citation in section when examining importance of citations. The same average time of citation mention implies method of measuring academic influence of a paper or author by total cited times can also be used in evaluation of citation in academic monographs.

Compared with academic papers, length of citation context in academic monographs is closer to the length of citation context, therefore, we are supposed to pay more attention when we make a syntactic analysis between citation context and citation content of
academic monographs. And we will explore specific reason of the difference in length of citation between academic monographs and academic papers.

**Conclusion and Future Works**

Nowadays, with the opening of full text database, more and more scholars begin to use full text of academic papers to explore citation content. On this basis, this article makes a comparative analysis of citation content in 39 academic monographs and 7,756 academic papers. The result indicates that there are obvious differences in location and length of citation between the AMs and APs, while citation mention of the two is familiar.

This paper is only a preliminary research, which still exists limitations. First of all, there is a gap between the data scale, since few full-text database of academic monographs are open to public. But our study has shown that the citation content of academic monographs has their own unique characteristics. Therefore, if full-text database was available, we can carry out a broader and deeper research on citation of academic monographs. Secondly, the AMs and APs studied in this paper are from two major disciplines. In the future, more disciplines and fields would be investigated using their monographs and papers. Finally, this article just carries out a syntactic research with some simple methods. We would extend our research from syntactic level to the semantic level, with help of the citation content and its context to explore the citation motivation and other semantic research work.

**Acknowledgments**

This work is supported by Major Projects of National Social Science Fund (No.16ZAD224), Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (No. MJUKF201704) and Qing Lan Project.

**References**


Guo, Z., Ding, Y. & Milojevi. (2013). Citation Content Analysis (CCA): A Framework for Syntactic and Semantic Analysis of Citation Content. *Journal of the Association for Information Science and Technology*, 64, 1490–1503


Lin, C. S., Chen, Y. F., & Chang, C. Y. (2013). Citation Functions in Social Sciences and Humanities - Preliminary Results from a Citation Context Analysis of Taiwan’s History Research Journals. *ASIST*.


Modeling Study of Knowledge Diffusion in Scientific Collaboration Networks based on Differential Dynamics

Yue Zenghui1, Xu Haiyun2, Yuan Guoting3, Pang Hongshen4, Dong Kun5

1 yzh66123@126.com
2 xuhai@clas.ac.cn
3 guotingyuan@hotmail.com
4 phs@szu.edu.cn
5 dongkun@mail.las.ac.cn

School of Medical Information Engineering, Jining Medical University, Rizhao (China)
Chengdu Documentation and Information Center, Chinese Academy of Sciences, Chengdu (China)
School of Foreign Languages, Jining Medical University, Rizhao (China)
Library, Shenzhen University, Shenzhen (China)
Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou (China)
Chengdu Documentation and Information Center, Chinese Academy of Sciences, Chengdu (China)

Abstract
Knowledge diffusion based on scientific collaboration is similar to disease propagation through actual contact. Inspired by the disease-spreading model in complex networks, this study classifies the states of research entities during the process of knowledge diffusion in scientific collaboration into four categories. Research entities can transform from one state to another with a certain probability, which results in the evolution rules of knowledge diffusion in scientific collaboration networks. The knowledge diffusion model of differential dynamics in scientific collaboration of non-uniformity networks is formed, and the relationship between the degree distribution and evolution of knowledge diffusion is further discussed, to reveal the dynamic mechanics of knowledge diffusion in scientific collaboration networks. Finally, an empirical analysis is conducted on knowledge diffusion in an institutional scientific collaboration network by taking the graphene field as an example. The results show that the state evolution of research entities in the knowledge diffusion process of scientific collaboration networks is affected not only by the evolution states of adjacent research entities with whom they have certain collaboration relationships, but also by the structural attributes and degree distributions of scientific collaboration networks. The evolution of knowledge diffusion in scientific collaboration entities with different degrees also shows different trends.

Conference Topic
Knowledge discovery and data mining

Introduction
Knowledge plays a vital role in economic growth, which is generally acknowledged by endogenous growth economists (Kline & Rosenberg, 1986; Lucas, 1988; Romer, 1990). The power of knowledge depends on whether the knowledge is diffused and on the diffusion depth and breadth, in addition to the value of the knowledge (Bacon, 1908). When knowledge is diffused, information and experiences gained by both sides would increase linearly; if the knowledge is further diffused with constant feedback and extension of the problems concerned, the information and experience would even increase geometrically (Quinn, Anderson, & Finkelstein, 1908). The knowledge societies are characterized by the proliferation of knowledge-intensive communities, specialized in knowledge production and reproduction, knowledge acquisition and exchange, and the use of information technologies (David & Foray, 2002; Yan, 2016). The production and creation of knowledge are not dependent on a single isolated entity; instead, knowledge is diffused, exchanged, and circulated among various entities (Crane, 1972). The competitiveness and the potential competitiveness of institutions are embodied in the acquisition and mastery of knowledge and innovative capacity (Drucker, 1999). Effective diffusion of knowledge can better promote their competitiveness and research level, and make optimal use of knowledge. Knowledge diffusion is the link between knowledge acquisition and knowledge application. It has become an important subject of common concern to realize more effective diffusion and management.
of knowledge (Cronin, 1982; Rogers, 2003; Chen & Hicks, 2004; Lambiotte & Panzarasa, 2009).

Many efforts have been made to improve the understanding of knowledge diffusion in various networks. According to the connection strength of network members, Granovetter (1973) proposed a weak ties theory of social networks, emphasizing the significance of "connection" between network members in knowledge and information diffusion. Cowan and Jonard (2004) and Kim and Park (2009) compared knowledge diffusion in regular, random, and small-world networks, and found that the small-world network is the most efficient structure to diffuse knowledge. Tang, Xi, and Ma (2006), and Lin and Li (2010) argued that the scale-free structure is more effective for knowledge diffusion.

Within the context of the growing complexity of research, collaboration has been considered one of the most crucial and common phenomena in the science community (Persson, Glänzel & Danell, 2004; Wuchty, Jones, & Uzzi, 2007; Adams, 2013; Chung, Kwon, & Lee, 2016). The process of scientific collaboration is also accompanied by the diffusing, sharing, and exchanging of both explicit and tacit knowledge among scientific research entities (Singh, 2005). Collective knowledge production and diffusion processes in science and technology have captured the attention of sociology of knowledge scholars throughout history (Mannheim, 1968; Kuhn, 1970; Scheler, 1980).

A scientific collaboration network, especially the co-authorship network formed by scientists in a deliberate and cautious way, is a structured form of knowledge exchanging and sharing among collaborators. The autonomous and self-organizing nature of scientific practices in knowledge creation and diffusion determines that the scientific collaboration network is the most appropriate mode of knowledge transmission and diffusion (Autant-Bernard, Mairesse & Massard, 2007; Ozel, 2010; Yang, Hu, & Liu, 2015).

Knowledge diffusion based on scientific collaboration is similar to disease propagation. In the process of scientific collaboration, knowledge exchange and diffusion can take place by the social collaboration connection among research entities including individuals, institutions, regions, countries, and so forth, while disease is usually propagated among organisms through air, food, contact, matrix, blood, and so forth (Anderson & May, 1991). From the perspective of information theory, both knowledge diffusion and disease propagation are composed of four elements that are the same in essence: information, information source, information channel, and information sink (Shannon, 1942).

The modeling of knowledge diffusion in scientific collaboration networks can help visualize the knowledge diffusion process through scientific collaboration, reveal the dynamics mechanism of knowledge diffusion in scientific collaboration networks, and realize more effective management and regulation of knowledge diffusion.

**Literature Review**

Studies on evolution mechanisms and the law of dynamics of knowledge diffusion are conducted from the perspective of model building, the current quantity of which is relatively small, with a majority of the research being qualitative. Usually, well-developed models from such fields as epidemiology, complexity science, and sociology are referenced. The knowledge diffusion process is accompanied by the generation and evolution of the knowledge diffusion network, and the construction of the knowledge diffusion model requires a tangible or intangible carrier network for knowledge diffusion. Currently, most scholars regard co-authorship (Eslami, Ebadi, & Schiffauerova, 2013) and literature citation relationships (Tsay, 2015; Zhu & Yan, 2015) as the paths to knowledge diffusion, and explore abstract representations of the knowledge diffusion process. The present knowledge diffusion models include the citation path model (Lu & Liu, 2013; Yu, Lu, Liu, & Zhou, 2014; Yan, 2014), epidemiological model (Bettencourt, Cinron-Arias, Kaiser, & Castillo-Chávez, 2006;
Epidemiological models focus on the transmission of different traits among certain populations; such traits can be transmitted diseases, knowledge, behaviors, or innovative ideas (Yan, 2014). An entity can be classified into one of the basic classes in epidemiological models: the susceptible class (S), the exposed class (E), the infected class (I), the skeptical class (Z), and the recovered class (R) (Hethcote, 2000). In the initial period of the diffusion of a good idea most of the population will be in the susceptible class (S), with a few entities in the exposed class (E)—having been in contact with the idea while not diffused it—and a small number of infected (I) manifesting it. In addition, there may be competing and mutually exclusive ideas (e.g., where susceptibles are turned off from the idea and become skeptics or idea stiflers, represented by the class Z). Furthermore, entities may recover or become immune (R), and not manifest the idea again (Bettencourt, Cinron-Arias, Kaiser, & Castillo-Chávez, 2006). Scholars can choose these classes based on their specific research questions.

In consideration of the similarity between knowledge diffusion based on scientific collaboration and disease propagation via actual contact, and inspired by the disease-spreading model in complex networks, the paper classifies research entities in the process of knowledge diffusion in scientific collaboration into potential knowledge recipients, potential knowledge diffusion entities, knowledge diffusion entities, and knowledge immunes. The four classifications of research entities can transform from one to another with a certain probability ($\alpha$, $\beta$, $\omega$, and $\gamma$, respectively). Then the evolution rules of knowledge diffusion in scientific collaboration networks are made. Furthermore, the knowledge diffusion model of differential dynamics in scientific collaboration of non-uniformity networks is formed, and then the relationship between the degree distributions and evolution of knowledge diffusion is further discussed, to reveal the dynamic mechanics of knowledge diffusion in scientific collaboration networks. Finally, an empirical analysis is conducted on knowledge diffusion in an institutional scientific collaboration network by taking the graphene field as an example.

**Theoretical Model**

*Description of knowledge diffusion process in scientific collaboration networks*

Many studies of the behavior dynamics mechanism of propagation and diffusion are based on the classical SIR epidemic model (Liu & Zhang, 2013; Su, Huang, & Zhao, 2015; Woo & Chen, 2016), in which diffusion states are defined to be the susceptible state (S), infective state (I), and recovered/immune state (R). During the actual evolution process of knowledge diffusion in scientific collaboration, however, there are two special phenomena: (1) When receiving certain knowledge, a research entity tends not to spread the knowledge immediately. Instead, there exists a certain period of time in which it digests the knowledge. After the research entity absorbs and converts the knowledge into intrinsic knowledge, it will spread the knowledge to others. (2) Under the influence of internal factors (such as the transition of research direction, demand, and interests) and external factors (such as the macro-control of the government), a research entity may not diffuse knowledge anymore. However, this kind of immunity is temporary for some research entities, because they will accept and diffuse the knowledge again in the future. Therefore, this paper argues that the SEIRS epidemic model, which has a certain latent state (E) and feedback mechanism (R to S), is more consistent with the nature of the knowledge diffusion process in scientific collaboration, and is more suitable
to function as the basic model of differential dynamics for knowledge diffusion in scientific collaboration networks.

**States of research entities in the knowledge diffusion process of scientific collaboration networks**

For the knowledge exchange and diffusion process among research entities, we define that research entities in different states of knowledge diffusion in scientific collaboration networks fall into four categories: potential knowledge recipients (S), potential knowledge diffusion entities (E), knowledge diffusion entities (I), and knowledge immunes (R).

1. Potential knowledge recipients (S): research entities who have not known the knowledge or have known but not acquired it yet.
2. Potential knowledge diffusion entities (E): research entities who have acquired the knowledge but not diffused it yet.
3. Knowledge diffusion entities (I): research entities who have mastered the knowledge and are diffusing it to the potential knowledge recipients.
4. Knowledge immunes (R): research entities who have acquired the knowledge but are immune to it now. They have lost interest in the knowledge and will not continue to diffuse it.

**Process of knowledge diffusion in scientific collaboration networks**

In the initial phase, there are only small numbers of knowledge diffusion entities in the network, while others are potential knowledge recipients. The number of potential knowledge diffusion entities and knowledge immunes is zero. As time goes on, knowledge begins to be diffused in the following ways:

1. When knowledge diffusion entities transmit knowledge to potential knowledge recipients, the recipients begin to accept it with a certain probability (α) and become potential knowledge diffusion entities.
2. If potential knowledge diffusion entities are interested in the knowledge, they will continue to diffuse it with a certain probability (β) and become new knowledge diffusion entities.
3. Knowledge diffusion entities lose interest in the knowledge with certain probability (ω) and turn into knowledge immunes.
4. Knowledge immunes become knowledge recipients again with a certain probability (γ) and take in the knowledge to which they have been immune.

**Evolution rules of knowledge diffusion in scientific collaboration networks**

At a certain time in the knowledge diffusion process in scientific collaboration networks, which could be marked as \( t \), a research entity can only be in one of the four states mentioned above. At the time node \( t \), we can define the proportion of scientific collaboration entities that are in a certain knowledge diffusion state in the whole system as follows.

1. \( s(t) \): the proportion of potential knowledge recipients to all research entities in different states of knowledge diffusion at time \( t \).
2. \( e(t) \): the proportion of potential knowledge diffusion entities to all research entities in different states of knowledge diffusion at time \( t \).
3. \( i(t) \): the proportion of knowledge diffusion entities to all research entities in different states of knowledge diffusion at time \( t \).
4. \( r(t) \): the proportion of knowledge immunes to all research entities in different states of knowledge diffusion at time \( t \).

Here, \( s(t)+e(t)+i(t)+r(t)=1 \).

According to the above description of the knowledge diffusion process in scientific collaboration networks, a schematic paradigm of the dynamic state evolution rules of research entities in knowledge diffusion is presented as follows (Figure 1).
Figure 1. State transformational rules of research entities’ knowledge diffusion of scientific collaboration.

When potential knowledge recipients cooperate with knowledge diffusion entities in the process of scientific collaboration, they could become potential knowledge diffusion entities with probability $\alpha$; potential knowledge diffusion entities become knowledge diffusion entities with probability $\beta$; knowledge diffusion entities turn into knowledge immunes with probability $\omega$; and knowledge immunes generate feedback with probability $\gamma$, and become potential knowledge recipients.

Modeling derivation of knowledge diffusion in scientific collaboration networks

A scientific collaboration network is a kind of complex networks with obvious scale-free features. Its degree distribution accords with the power-law distribution. The connections between nodes are unevenly distributed, which means that only a minority of nodes in the network have many links, while most nodes have few links (Barabási & Albert, 1999). Research entities of different degrees play different roles in the process of knowledge diffusion. According to the evolution rule of knowledge diffusion mentioned above, the knowledge diffusion model of differential dynamics in scientific collaboration of non-uniformity networks is formed by applying the mean field theory on the basis of the classical disease-spreading equation of SEIRS. The derivation process is deduced as follows.

In accordance with the evolution rule shown in Figure 1, considering the non-uniformity distribution characteristics of nodes, the system dynamics equations for each point $(k)$ can be established from the perspective of mean field theory as follows:

\[
\begin{cases}
\frac{ds_k(t)}{dt} = -\alpha k N(k)s_k(t)\theta_k(t) + \gamma r_k(t) \\
\frac{de_k(t)}{dt} = \alpha k N(k)s_k(t)\theta_k(t) - \beta e_k(t) \\
\frac{di_k(t)}{dt} = \beta e_k(t) - \omega i_k(t) \\
\frac{dr_k(t)}{dt} = \omega i_k(t) - \gamma r_k(t)
\end{cases}
\] (1)

In the equation set above, $t$ is the time step; $k$ is the node degree of research entities; $N(k)$ stands for the number of research entities with degree $k$; $s_k(t)$, $e_k(t)$, $i_k(t)$, and $r_k(t)$ indicate the proportion of research entities with degree $k$ at time $t$, and their knowledge diffusion states are $S$, $E$, $I$, and $R$, respectively; and $\theta_k(t)$ represents the probability that a random edge will connect with any knowledge diffusion entity at time $t$ (namely, the infection probability for a research entity through a linked edge with other entities).

For a scientific collaboration network in which node degrees are correlative, $\theta_k(t)$ can be expressed as (Xia, Liu, Chen, & Yuan, 2008):

\[
\theta_k(t) = \sum_{k'} P(k' | k)i_{k'}(t)
\] (2)

where $P(k' | k)$ stands for the probability that an edge will stretch from a node with degree $k$ to a node with degree $k'$.

If the node degrees are not correlative, then
\[
\theta_k(t) = \sum_{k'} P(k' \mid k) i_{k'}(t) = \sum_{k'} \frac{k'P(k')i_{k'}(t)}{<k>}
\]

where \(<k>\) is the average degree of the scientific collaboration network; \(P(k')\) indicates the probability that the research entity with degree \(k'\) will collaborate with an entity whose degree is \(k\); and \(i_{k'}(t)\) represents the proportion of research entities with degree \(k'\) whose knowledge diffusion state is \(I\) at time \(t\).

By putting Formula 2 and Formula 3 into Equation Set 1, respectively, the knowledge diffusion models of differential dynamics in scientific collaboration networks with both correlated and uncorrelated node degrees will be formed, eventually, which are represented respectively as follows.

(1) The knowledge diffusion model of differential dynamics in scientific collaboration networks whose node degrees are correlative:

\[
\begin{align*}
\frac{ds_{s}(t)}{dt} &= -\alpha k N(k) s_{s}(t) \sum_{k'} P(k' \mid k) i_{k'}(t) + \gamma r_{s}(t) \\
\frac{de_{s}(t)}{dt} &= \alpha k N(k) s_{s}(t) \sum_{k'} P(k' \mid k) i_{k'}(t) - \beta e_{s}(t) \\
\frac{di_{s}(t)}{dt} &= \beta e_{s}(t) - \omega i_{s}(t) \\
\frac{dr_{s}(t)}{dt} &= \omega i_{s}(t) - \gamma r_{s}(t)
\end{align*}
\]

(2) The knowledge diffusion model of differential dynamics in scientific collaboration networks whose node degrees are uncorrelated:

\[
\begin{align*}
\frac{ds_{s}(t)}{dt} &= -\alpha k N(k) s_{s}(t) \sum_{k'} \frac{k'P(k')i_{k'}(t)}{<k>} + \gamma r_{s}(t) \\
\frac{de_{s}(t)}{dt} &= \alpha k N(k) s_{s}(t) \sum_{k'} \frac{k'P(k')i_{k'}(t)}{<k>} - \beta e_{s}(t) \\
\frac{di_{s}(t)}{dt} &= \beta e_{s}(t) - \omega i_{s}(t) \\
\frac{dr_{s}(t)}{dt} &= \omega i_{s}(t) - \gamma r_{s}(t)
\end{align*}
\]

**Empirical Research**

Graphene has drawn worldwide research attention because of its unique structure and the excellent characteristics of electricity, mechanics, optics, chemistry, and thermodynamics, becoming one of the hottest research subjects in the fields of physics, chemistry, and material science (Ma, Wan, & Feng, 2012). Scientific collaboration in the field is very common and frequent, and the transmission and communication of knowledge are very active. Carbon nanotubes are allotropes of carbon with a cylindrical nanostructure and one-dimensional quantum materials with a special structure. Graphene and carbon nanotubes are complementary in structure and capability, and there are also marked resemblances between them in research methodology (Baughman, Zakhidov, & De Heer, 2002). Studies of them present an overlapping, mutually permeating, and inseparable trend. Therefore, the knowledge point "carbon nanotubes" in the field of graphene is selected as the research object to form and verify the knowledge diffusion model of differential dynamics that simulates the knowledge diffusing process of carbon nanotubes through the institutional scientific collaboration network of graphene, and the research results are further analyzed and explained. Because the knowledge point "carbon nanotubes" in the field of graphene emerged
in 1993, the research period of this model is set from 1993 to 2012.

Data collection

According to the following rules, the data was retrieved from the Web of Science (Table 1).

Table 1. Rules of data retrieval.

<table>
<thead>
<tr>
<th>Retrieval strategy</th>
<th>TS=(graphen* or &quot;single layer graphit*&quot; or &quot;monolayer graphit*&quot;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>SCI-EXPANDED, SSCI, CPCI-S (Conference Proceedings Citation Index – Science)</td>
</tr>
<tr>
<td>Document type</td>
<td>Articles, Proceedings Paper</td>
</tr>
<tr>
<td>Period</td>
<td>1990-2012</td>
</tr>
</tbody>
</table>

By applying the above retrieval rules, 23,458 primary pieces of literature were obtained. Then the data was imported into the Thomson Data Analyzer (TDA) and data cleaning was conducted. The institutional scientific collaboration network matrices of graphene and institutional knowledge diffusion network matrices over the years were ultimately formed. Considering the updating of the network database, the data was collected on June 7, 2014 to maintain data consistency.

Degree correlation of scientific collaboration networks

The Pearson correlation score between the node degree k and the average degree of its adjacent nodes is -0.402. A significant correlation is not observed between them. Therefore, we ultimately select Model 5 as the theoretical model of this study.

State evolution of knowledge diffusion entities

In this paper, we define research entities in the state of knowledge diffusion entities at a certain time as the institutions that publish papers containing the knowledge point "carbon nanotubes" in the graphene field at that time. The total number of members in the institutional scientific collaboration network is 2,595. The quantity and proportion of knowledge diffusion entities from 1993 to 2012 are shown in Table 2.

Table 2. Quantity and proportion of knowledge diffusion entities during 1993–2012.

<table>
<thead>
<tr>
<th>Year</th>
<th>Quantity</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1993</td>
<td>3</td>
<td>0.001156</td>
</tr>
<tr>
<td>1994</td>
<td>14</td>
<td>0.005395</td>
</tr>
<tr>
<td>1995</td>
<td>8</td>
<td>0.003083</td>
</tr>
<tr>
<td>1996</td>
<td>16</td>
<td>0.006166</td>
</tr>
<tr>
<td>1997</td>
<td>22</td>
<td>0.008478</td>
</tr>
<tr>
<td>1998</td>
<td>45</td>
<td>0.017341</td>
</tr>
<tr>
<td>1999</td>
<td>42</td>
<td>0.016185</td>
</tr>
<tr>
<td>2000</td>
<td>58</td>
<td>0.022351</td>
</tr>
<tr>
<td>2001</td>
<td>67</td>
<td>0.025819</td>
</tr>
<tr>
<td>2002</td>
<td>90</td>
<td>0.034682</td>
</tr>
<tr>
<td>2003</td>
<td>93</td>
<td>0.035838</td>
</tr>
<tr>
<td>2004</td>
<td>113</td>
<td>0.043545</td>
</tr>
<tr>
<td>2005</td>
<td>151</td>
<td>0.058189</td>
</tr>
<tr>
<td>2006</td>
<td>171</td>
<td>0.065896</td>
</tr>
<tr>
<td>2007</td>
<td>230</td>
<td>0.088632</td>
</tr>
<tr>
<td>2008</td>
<td>294</td>
<td>0.113295</td>
</tr>
<tr>
<td>2009</td>
<td>401</td>
<td>0.154528</td>
</tr>
<tr>
<td>2010</td>
<td>577</td>
<td>0.222351</td>
</tr>
<tr>
<td>2011</td>
<td>790</td>
<td>0.304432</td>
</tr>
<tr>
<td>2012</td>
<td>968</td>
<td>0.373025</td>
</tr>
</tbody>
</table>
It can be seen from the table that the proportion of institutions in the state of knowledge diffusion entities (in other words, institutions that are diffusing knowledge), is growing continually (from 0.1156% to 37.3025%). The growth rate is slow from 1993 to 2006, and gradually speeds up after 2006. In 1995 and 1999, though, the proportion declines. The knowledge diffusion of carbon nanotubes is in the embryonic stage from 1993 to 2006, and enters its development stage from 2007 to 2012.

Model construction and validation

In the original state (i.e., in 1993), there were three knowledge diffusion entities of carbon nanotubes in the network (i.e., Massachusetts Institute of Technology; University of California, Los Angeles; and Drexel University). The others were all potential knowledge recipients. The numbers of both potential knowledge diffusion entities and knowledge immunes were zero.

Parameter determination

According to the distributing characteristics of the theoretical and practical data values (between zero and one), the Kolmogorov–Smirnov statistic and Likelihood function are employed in this paper for parameter estimation. On the basis of Model 5, we simulate the theoretical model using MATLAB, adjust parameters ($\alpha$, $\beta$, $\omega$, and $\gamma$) with a step length of 0.1, and calculate K-S statistic and MLE (L) between the output values of the model with different parameters and actual values from 1993 to 2009, respectively, to determine the optimal parameters of the fitting model. The set of parameters that result in the minimum values for both K-S and L is the optimal parameter set for the fitting model at the 0.05 significance level (Table 3).

Table 3. Optimal parameter set of fitting model and its K-S and L values.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\omega$</th>
<th>$\gamma$</th>
<th>K-S</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.107895</td>
<td>3.670283</td>
</tr>
</tbody>
</table>

Model building

Evolution curves of the theoretical and actual proportions of knowledge diffusion entities in the optimal fitting model are shown in Figure 2.

Figure 2. Evolution curves of theoretical and actual proportions of knowledge diffusion entities with the optimal fitting parameters.

Therefore, the knowledge diffusion model of differential dynamics simulating the knowledge diffusion of carbon nanotubes through the institutional scientific collaboration network of graphene can be described as follows.
The model illustrates that the state evolution of research entities in the knowledge diffusion process of scientific collaboration networks is affected not only by the evolution states of adjacent research entities with whom they have certain collaboration relationships in knowledge diffusion, but also by the structural attributes and degree distributions of scientific collaboration networks. The state change of institutions with different node degrees (k) is influenced by factors such as their own degrees, the number of nodes (N(k)), and the connection probability of states.

Model verification

In accordance with Model 6, the proportions of institutions that are diffusing knowledge from 2010 to 2012 are predicted by the iteration of time steps (Table 4).

### Table 4. Predicted value of the proportion of knowledge diffusion entities from 2010 to 2012.

<table>
<thead>
<tr>
<th>Year</th>
<th>Predicted value</th>
<th>Actual value</th>
<th>Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>0.217252</td>
<td>0.222351</td>
<td>-2.293%</td>
</tr>
<tr>
<td>2011</td>
<td>0.300011</td>
<td>0.304432</td>
<td>-1.452%</td>
</tr>
<tr>
<td>2012</td>
<td>0.377568</td>
<td>0.373025</td>
<td>1.218%</td>
</tr>
</tbody>
</table>

The table shows that the deviations between the theoretical value and actual value of the proportion of knowledge diffusion entities predicted from 2010 to 2012 are -2.293%, -1.452%, and 1.218%, respectively. This is an insignificant discrepancy, and the closeness between the values illustrates the validity of the model.

### Results

**Evolution analysis of knowledge diffusion in scientific collaboration networks**

From 1993 to 2012, the state evolution of research entities diffusing the knowledge of carbon nanotubes in the institutional scientific collaboration network of graphene is shown in Figure 3.

![Figure 3. Evolution curves of research entities in different knowledge diffusion states in the scientific collaboration network.](image-url)
From the perspective of the overall evolution trend of knowledge diffusion, the proportion of potential knowledge recipients in the institutional scientific collaboration network shows an incessant drop with the passage of time. However, the proportion of knowledge diffusion entities and knowledge immunes rises constantly. Moreover, the proportion of potential knowledge diffusion entities shows only a slight growth trend. From the perspective of evolution tempo (changing speed of proportion) of knowledge diffusion, the proportion of potential knowledge recipients, knowledge diffusion entities, and knowledge immunes changes slowly at first and then speeds up rapidly, while the change of proportion of potential knowledge diffusion entities has shown a slight growth tendency. From the perspective of evolution acceleration (change rate of velocity) of knowledge diffusion, the change rate of potential knowledge recipients ranks first, followed by that of knowledge diffusion entities and knowledge immunes, and that of potential knowledge diffusion entities ranks last.

In addition, it can be judged from the evolution trend of state proportion of knowledge diffusion entities that the knowledge diffusion of carbon nanotubes in the institutional scientific collaboration network of graphene is in its developing stage, having not reached the saturation point. The diffusion scale of the knowledge will continue to expand as time goes on.

Relationship between node degree distribution and knowledge diffusion evolution in scientific collaboration networks

To reveal the impact of node degrees on knowledge diffusion, the relationship between node degree distribution and knowledge diffusion evolution in scientific collaboration networks is further analyzed and explained.

(1) Node degrees and potential knowledge recipients in scientific collaboration networks
The evolution of the proportion of nodes with variable degrees k in the state of potential knowledge recipients is shown in Figure 4.

![Figure 4. Evolution curves of proportion of nodes with variable degrees k in the state of potential knowledge recipients.](image)

We can see from the above evolution curves that, overall, the proportion of potential knowledge recipients shows a downward trend, with the descending rates slowing down gradually. The proportion of scientific research entities with degree k=2 remains at a low level. Simultaneously, the proportion of potential knowledge recipients at the nodes with larger degrees is higher.

(2) Node degrees and potential knowledge diffusion entities in scientific collaboration networks
The evolution of the proportion of nodes with variable degrees k in the state of potential
knowledge diffusion entities is shown in Figure 5.

![Figure 5](image)

**Figure 5. Evolution curves of proportion of nodes with variable degrees k in the state of potential knowledge diffusion entities.**

As shown in Figure 8, generally speaking, the proportion of potential knowledge diffusion entities is in decline, and the declining rates slow down by degrees. The proportion in research entities with degree k=2, however, is rising slowly, and the proportion with degree k=1 rises rapidly first, followed by a sharp drop, and finally becomes steady.

(3) Node degrees and knowledge diffusion entities in scientific collaboration networks

The evolution of the proportion of nodes with variable degrees k in the state of knowledge diffusion entities is illustrated in Figure 6.

![Figure 6](image)

**Figure 6. Evolution curves of proportion of nodes with variable degrees k in the state of knowledge diffusion entities.**

As shown in Figure 9, the overall trend of the proportion of knowledge diffusion entities rises rapidly at first and then descends slowly. The proportion in research entities with degree k=2, however, shows a gradual falling trend. Regarding scientific research entities with node degree k before reaching the highest proportion of the knowledge diffusion entities, the larger the degree is, the lower the proportion will be; after the highest proportion of the knowledge diffusion entities, the larger the degree is, the higher the proportion will be.

(4) Node degrees and knowledge immunes in scientific collaboration networks

The evolution of the proportion of nodes with variable degrees k in the state of knowledge immunes is shown in Figure 7.
Generally, the proportion of knowledge immunes with variable degrees rises gradually. Simultaneously, the larger the degree is, the lower the proportion of knowledge immunes will be.

Additionally, in light of the assumption of the model that knowledge diffuses with the help of scientific collaboration, the node with degree $k=0$ will not engage in the knowledge diffusion process because it does not participate in scientific collaboration. Therefore, it will remain in the state of potential knowledge diffusion entities. Once there is scientific collaboration, the node degree will change and be further involved in the recurrent state of knowledge diffusion.

**Discussion**

Inspired by the disease-spreading model in complex networks, the paper builds the knowledge diffusion model of differential dynamics in scientific collaboration of non-uniformity networks, and conducts an empirical analysis on the model. The research has shown that:

1. The hypothesis that knowledge can be diffused with the aid of scientific collaboration is justified, and the knowledge diffusion model constructed in this paper is rational and reasonable.
2. The state evolution of research entities in the knowledge diffusion process of scientific collaboration networks is affected not only by the evolution states of adjacent research entities with whom they have certain collaboration relationships, but also by the structural attributes and degree distributions of scientific collaboration networks. For institutions with different node degrees, the change of states is jointly influenced by their own degrees, the number of nodes, and the connection probability of states.
3. When the state transition probability meets the conditions that $\alpha=0.3$, $\beta=1$, $\omega=0.1$, and $\gamma=0.1$, the knowledge diffusion model of differential dynamics can almost accurately simulate the knowledge diffusion process of carbon nanotubes in the institutional scientific collaboration network of graphene.
4. In the evolution process of knowledge diffusion in the scientific collaboration network, the proportion of potential knowledge recipients declines constantly, but the proportion of knowledge diffusion entities and knowledge immunes increases continuously. Meanwhile, the proportion of potential knowledge diffusion entities shows only a slight growth trend.
5. The knowledge diffusion evolution of scientific collaboration entities with different degrees also shows different trends.

Based on the previous research results, policy makers can regulate the structure, degree distributions, and transition probability of states in scientific collaboration networks through various means to achieve different evolution effects of knowledge diffusion, and can control
the knowledge diffusion process in scientific collaboration by promoting or inhibiting it. For example, they could expand the cooperation breadth of research institutions to further promote the absorption and mastery of knowledge and accelerate diffusion of knowledge with measures such as providing platforms for academic communication, consummating incentive mechanisms for scientific research innovation, etc.

Acknowledgements
This research is supported by Shandong Provincial Natural Science Foundation, China (Grant No. ZR2015GL015), Doctoral Scientific Research Foundation of Jining Medical University (Grant No. JY2015BS14), Science and Technology Development Project of Jining (Grant No. 2014jnyf15), and Youth Foundation of Jining Medical University (Grant No. JYQ14RW17).

References


Research of cross-discipline based on Web-of-Science Categories

Lin Yuan  Liu Haifeng  Sun Xiaoling  Ding Kun  Li Luying

dingk@dlut.edu.cn
Dalian University of Technology, Dalian (China )

Abstract
There were 11398 journals were included in Journal Citation Report from Web of Science in 2015, including about 44.47 percent of journals included by multiple disciplines, which indicates interdisciplinary is a common phenomenon. The number of cross-journals among different disciplines reveals the role of journals in the interdisciplinary, which can be used to analysis interdisciplinary. This paper proposes a Cross-Discipline Vector Space Model for quantitative analysis of the degree of interdisciplinary. Through calculating the relevance among discipline, we obtain the pattern of interdisciplinary, which could provide scientific advice to promote the development of disciplines and scientific research.

Keywords: Web of Science Categories, Cross-Discipline, Vector Space Model, Scientometrics.

Conference Topic
Scientometrics, Webometrics and Knowledgometrics: The research about data science and metrology; Journals, databases and electronic publications; Methods and techniques; Knowledge discovery and data mining

Introduction
With the development of scientific research activities, scientists engaged in the activities of scientific research from single field to broader fields. The inner cross-discipline also influences the scientific research activities. The development of discipline emerges in two trends. One is the overall subject gradually divided into a finer professional field, the other one is the discipline overlapping forming a new discipline, which is due to the progress of science and technology and the demand for society to solve practice requirement, different discipline fields become more and more close. For example, Artificial Intelligent is a brand-new discipline field applying computer technology and biological neural network technology comprehensively. Therefore, how to quantify the degree of correlation between disciplines becomes more and more important. Revealing the development of cross-discipline will provide reference for the advancement of scientific research in the future.

Cross-Discipline is a comprehensive scientific activity which is born to meet the needs of social development and the development of discipline itself. Cross-discipline is the prerequisite for formatting interdisciplinary, and the interdisciplinary is the result of Cross-discipline(Xu, Yi & Guo, 2015, PORTER& RAFOLS,2009). Cross-Discipline is the new growth point of science and the new frontier of science. Scientific achievements are often originated in interdisciplinary field. Interdisciplinary sciences is a comprehensive and systematic knowledge system which is formed in the intersection of the external disciplines and the inter discipline of natural sciences, social sciences, humanities and other disciplines(Li, Zong&Xu,2013). The continuous development of interdisciplinary sciences has greatly promoted the progress of science. Therefore, Cross-Discipline research reflects the trend of science development. With the rapid development of discipline in science field and the strengthening of university cooperation, disciplines have broken the previous barriers and the disciplinary integration has gradually produced brand new disciplines.

The studies of Cross-Discipline are mainly divided into the following research methods: Researchers used Citation Analysis Methods to explore the characteristics of knowledge diffuse among different disciplines(Qiu,Cao,2012). Researchers used Cross-Analysis to analysis literatures based on target literature, references and cited literature(Huang,
According to the practical research experience of Cross-Discipline, Weiying Jin used the method of literature and system theory to recognize Cross-Discipline (Jing, 2005). Xiaoling Sun used cross-field authors to map the evolution of scientific fields (Sun, Ding & Lin, 2016).

In this paper, we use the statistical method to analyze 11398 core journals which were included in SCIE and SSCI. In order to ind out the degree of correlation about Cross-Discipline and reveal the associated rules about Cross-Discipline, we proposed a Cross-Discipline Vector Space Model method to analyze the number of core journals crossed by disciplines. The degree of correlation between Cross-Discipline have a further influence on discipline research and national development strategy.

Materials and methods

Dataset

In this paper, we use ISI (Institute for Scientific Information) database platform of WOS (Web of Science) to study Cross-Discipline. Web of Science is the products based on Web platform by Thomson Reuters Company, and is the largest comprehensive academic information resource which covered the largest number of disciplines in the world. Web of Science includes 11398 high-quality core journals in 2015, which are the most influential journals and have been reviewed by peers around the world. We used these core journals included in the database for research. Web of Science divides all disciplines into 234 categories, and every year Web of Science will publish an edition of journal classification from 1997 to 2015. We use the latest discipline division of 2015 for analysis.

By logging into WOS database, and entering JCR (Journal Citation Reports) module, you can download all the journal information included by 234 categories in 2015.

Research Methods

SCI/SSCI journals are not always classified into a single JCR category, some journals are grouped into multiple JCR categories, which are called “Multi-assigned Journal” (Li & Chen, 2015). In this paper, we firstly use the statistical methods to compare the discipline of Multi-assigned Journal in 234 disciplines, and then use the Cross-Discipline Vector Space Model to calculate the degree of interdiscipline based on the number of Multi-assigned Journal in different disciplines. The Vector Space Model is the concept of continuous correlation sequence proposed by (Salton, Wong & Yang, 1975). The correlation measure method is widely used in information retrieval, but it rarely applied in discipline relevance measurement, so we select Vector Space Model to measure the degree of correlation between disciplines.

Results

Core journals included by disciplines

In scientific fields, We take the percentage the number of interdisciplinary journals in a discipline take up in all the core journals as the indicator to measure the interdisciplinary (Cheng & Liu, 2005). Figure 1 shows the Cross-Discipline percentage distribution of 234 disciplines.
Figure 1 shows that there are 46 disciplines whose percentage of multi-assigned journals is between 80% and 90%. While there are only 1 discipline whose percentage of multi-assigned journals is between 0% and 10%. Overall, most disciplines have more than 50% (a total of 178 disciplines, 76.1% of all disciplines) of multi-assigned journals, which is in a higher range.

From the data in Figure 1, it can be concluded that Cross-Discipline is a common phenomenon. Compared with the situation of Cross-Discipline in previous decade, Nowadays, the interdisciplinary integration of the disciplines on the basis of the original moves further forward, and the degree of cross-range is more far-reaching and more extensive span. It can be concluded that discipline development will enter a new era in the future, the barriers between disciplines will no longer exist and how to grasp the degree of interdisciplinary with an accurately method will be related to the development of scientific research activity and academic research.

The Cross-Discipline situation based on the core journals
Web of Science divided 234 disciplines and contain a total of 11398 journals. There are large number of Multi-assigned journals exist in 234 disciplines. The detail is showed in Figure 2. As we can see from Figure 2, the number of interdisciplinary journals with one Multi-assigned Journals are 6298, that is, there are 6298 journals that were only included in self-disciplines in Web of Science. There are 5100 journals belong to Multi-assigned journals, accounting for a total of 44.74%. The number of Multi-assigned journals was so large, which indicates that there is a large degree of Cross-Discipline between the disciplines. Specially, there are 8 journals included by 6 disciplines at the same time, the name of which are: 1. JOURNAL OF CHEMOMETRICS 2. ADVANCED MATERIALS 3. EUROPEAN JOURNAL OF CANCER CARE 4. CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS 5. Disability and Health Journal 6. NANO LETTERS 7. Small 8. ADVANCED FUNCTIONAL MATERIALS. There are two core journals belong to materials among eight kinds of journals, two core journals belong to nanotechnology, two core journals belong to medical and health science, one core journal in the field of chemistry.
and one core journal of computer science. It is found that there is large similarity between materials science and nanotechnology in the eight journals by analyzing the similarity of the disciplines later in this paper. The two disciplines not only have a large degree of Cross-Discipline between themselves, but also have Cross-Discipline with other disciplines. Meanwhile, there are 6298 core journals only included by one discipline, which have more pertinence and researchers will more convenient to get relative information by query those journals for research question of specific areas.

![Figure 2. The distribution of journals based on the number of interdisciplinary subjects](image)

**Cross-Discipline analysis between disciplines**

For further study, we process data of each Cross-Discipline in details. Figure 3 shows the result of Cross-Discipline situations.

![Figure 3. The discipline distribution based on the number of interdisciplinary subjects](image)

We found that the number of subjects in 234 categories without interdisciplinary subjects was zero. In other words, a journal that does not have a single subject belongs to itself. Meanwhile, from the distribution of Figure...
3 we can find that a total of 96 disciplines cross with four disciplines, and there are 202 disciplines have Cross-Discipline with at least three disciplines, accounting for 86.3% of the total number of disciplines. It is shown that discipline has not only intersected with one or two disciplines, but mostly intersects with more than three disciplines. The interdisciplinary phenomenon shows that there are close relationships between disciplines. By comparison, only 32 disciplines have fewer than three interdisciplinary disciplines. We could see that the disciplines have completely broken the academic barriers (Zhu, 2003), exploring the Cross-Discipline among disciplines will be a shortcut to advance the convergence of disciplines in the future.

At this point, we have demonstrated that there are very close relationships between disciplines and the Cross-Discipline has reached a certain height. How to quantify the degree of interdisciplinary will be related to the further development of disciplines.

Cross-Discipline Vector

In order to measure the correlation between disciplines, we propose a Cross-Discipline Vector Space Model for quantitative analysis of the degree of interdisciplinary. The motivation of core journals included in a discipline is the importance for the development of the discipline, regardless if the journal is the core journal for other disciplines or not. All the core journals included in the same discipline have equivalent contribution to the discipline, and they have the same property for the discipline. Based on the above viewpoint, we propose the Cross-Discipline Vector Space Model and calculate the similarity between disciplines. Because of the Vector Space Model can perfectly represent the data relevance and has been widely used to calculate the similarity between data in information retrieval, we use the Vector Space Model to calculate the similarity and based on the values of spatial Vector to determine the degree of Cross-Discipline.

Vector definition

We build model to explore the relevance of disciplines based on 234 disciplines. The basic vector of discipline research is constituted by the number of Multi-assigned journals between this discipline and other disciplines. In order to measure the degree of Cross-Discipline between A and B, we use C discipline as a bridge. We use the number of Multi-assigned journals between A and C as an index, and the number of Multi-assigned journals between B and C as another index, then the relevance between A discipline and B discipline is calculated using these two indexes as input for Cross-Discipline Vector Space Model. The construction of the vector is shown in the Table 1.

<table>
<thead>
<tr>
<th>Table</th>
<th>LAW</th>
<th>ENVIRONMENTAL STUDIES</th>
<th>MATHEMATICS APPLIED</th>
<th>.. LOGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECONOMICS</td>
<td>9</td>
<td>16</td>
<td>0</td>
<td>.. 0</td>
</tr>
<tr>
<td>MATHEMATICS</td>
<td>1</td>
<td>0</td>
<td>111</td>
<td>.. 12</td>
</tr>
</tbody>
</table>

The discipline in this paper is represented as a Vector with 234 dimensions, and the value of each dimension stands for the number of common core journals included by this discipline and other disciplines at the same time, and the corresponding discipline of every dimension is the same for all the disciplines. For example, in Table 1, ECONOMICS discipline is represented as a vector of 234 dimensions, and each dimension stands for a discipline of 234 disciplines. The dimension value is the number of core journals that intersected with other disciplines. If there are no common core journals between disciplines, the value of the dimension will be zero.
The calculation of vector relevance.

We use the cosine of the angle between the spatial vectors to calculate similarity between Cross-Discipline Vectors. The calculation method is shown as formula (1):

$$\text{Cosine}(x,y) = \frac{\sum_{j=1}^{t} c_{xj} c_{yj}}{\sqrt{\sum_{j=1}^{t} (c_{xj})^2} \sqrt{\sum_{j=1}^{t} (c_{yj})^2}}$$

(1)

where $x, y$ are two disciplines, $X, Y$ are discipline vectors, $c_{xj}$ is the number of Multi-assigned journals between $x$ discipline and $j$ discipline, $t$ is the number of disciplines, which is 234 there. $j$ is the number of disciplines which have included same journals with $x$ (including $x$ itself). The result is a positive number which is in the range $[0,1]$, and the result is 1 when the two Vectors are exactly equal. The larger the similarity between the vectors is, the higher the value of cosine method will be. In this paper, we can get the similarities with 234 disciplines based on the Multi-assigned journals, and we choose the disciplines with the values of similarity larger than 0.5 to show in Table 2.

### Table 2. Top 46 disciplines ranked by similarity.

<table>
<thead>
<tr>
<th>Rank</th>
<th>A discipline</th>
<th>B discipline</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HEALTH POLICY &amp; SERVICES</td>
<td>HEALTH CARE SCIENCES &amp; SERVICES</td>
<td>0.78</td>
</tr>
<tr>
<td>2</td>
<td>EDUCATION SPECIAL</td>
<td>REHABILITATION</td>
<td>0.75</td>
</tr>
<tr>
<td>3</td>
<td>MEDICAL ETHICS</td>
<td>ETHICS</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>ENGINEERING PETROLEUM</td>
<td>ENERGY &amp; FUELS</td>
<td>0.75</td>
</tr>
<tr>
<td>5</td>
<td>IMAGING SCIENCE &amp; PHOTOGRAPHIC TECHNOLOGY</td>
<td>REMOTE SENSING</td>
<td>0.75</td>
</tr>
<tr>
<td>6</td>
<td>PHYSICS NUCLEAR</td>
<td>PHYSICS PARTICLES &amp; FIELDS</td>
<td>0.73</td>
</tr>
<tr>
<td>7</td>
<td>NANOSCIENCE &amp; NANOTECHNOLOGY</td>
<td>MATERIALS SCIENCE MULTIDISCIPLINARY</td>
<td>0.73</td>
</tr>
<tr>
<td>8</td>
<td>GERONTOLOGY</td>
<td>GERIATRICS &amp; GERONTOLOGY</td>
<td>0.71</td>
</tr>
<tr>
<td>9</td>
<td>NANOSCIENCE &amp; NANOTECHNOLOGY</td>
<td>PHYSICS APPLIED</td>
<td>0.69</td>
</tr>
<tr>
<td>10</td>
<td>CONSTRUCTION &amp; BUILDING TECHNOLOGY</td>
<td>ENGINEERING CIVIL</td>
<td>0.68</td>
</tr>
<tr>
<td>11</td>
<td>MATHEMATICS APPLIED</td>
<td>MATHEMATICS</td>
<td>0.68</td>
</tr>
<tr>
<td>12</td>
<td>ALLERGY</td>
<td>IMMUNOLOGY</td>
<td>0.68</td>
</tr>
<tr>
<td>13</td>
<td>SOCIAL SCIENCES &amp; MATHEMATICAL METHODS</td>
<td>MATHEMATICS INTERDISCIPLINARY APPLICATIONS</td>
<td>0.67</td>
</tr>
<tr>
<td>14</td>
<td>AGRICULTURAL ECONOMICS &amp; POLICY</td>
<td>ECONOMICS</td>
<td>0.65</td>
</tr>
<tr>
<td>15</td>
<td>PSYCHOLOGY BIOLOGICAL</td>
<td>BEHAVIORAL SCIENCES</td>
<td>0.65</td>
</tr>
<tr>
<td>16</td>
<td>BIODIVERSITY CONSERVATION</td>
<td>ECOLOGY</td>
<td>0.65</td>
</tr>
<tr>
<td>17</td>
<td>PHYSICS APPLIED</td>
<td>MATERIALS SCIENCE MULTIDISCIPLINARY</td>
<td>0.64</td>
</tr>
</tbody>
</table>
From the perspective of journals, we analyze the similarities among 234 disciplines of Web of Science. The study found that there are 46 pairs of disciplines with similarity larger than 0.50 in
234 disciplines, and 8 pairs of disciplines with similarity larger than 0.7. The similarities between disciplines are shown in Table 2; The 8 pairs of disciplines are HEALTH POLICY & SERVICES and HEALTH CARE SCIENCES & SERVICES, EDUCATION SPECIAL and REHABILITATION, MEDICAL ETHICS and ETHICS and so on. The highest degree of similarity between disciplines is HEALTH POLICY & SERVICES and HEALTH CARE SCIENCES & SERVICES, the similarity of which is 0.78. From this result, we could see that these pairs of disciplines have very similar research topics compared to other disciplines, and these disciplines have completed a large degree of Cross-Discipline. There are total 82 disciplines with similarities that are more than 0.50, accounting for 35.04% of all discipline, indicating that the current development of disciplines show a rapidly cross trend, and nearly half of the literature and knowledge are interrelated. Plenty of same core journals are included in different disciplines, which indicate Cross-Discipline is a common phenomenon. Specially, we find that there are higher degrees of similarity among multiple disciplines. Such as NANOSCIENCE & NANOTECHNOLOGY, MATERIALS SCIENCE MULTIDISCIPLINARY and PHYSICS APPLIED. The similarity between NANOSCIENCE & NANOTECHNOLOGY and MATERIALS SCIENCE MULTIDISCIPLINARY is 0.73, and the similarity between NANOSCIENCE & NANOTECHNOLOGY and PHYSICS APPLIED is 0.69, the similarity between PHYSICS APPLIED and MATERIALS SCIENCE MULTIDISCIPLINARY is 0.64. In all disciplines, these three disciplines rank in the top 20 by similarity, indicating that the Cross-Discipline existence is not limited to two disciplines, but is forming a Cross-Discipline cluster over time, and the interdisciplinary phenomenon between discipline clusters indicate that disciplinary development has entered the stage of discipline cluster, and the Cross-Discipline as discipline clusters is evolving. In all disciplines of Cross-Discipline, there are 12 pairs of medical disciplines with similarity larger than 0.50, accounting for 26.08% of all 46 pairs, indicating that medical disciplines have better Cross-Discipline among all disciplines in recent years and medical discipline is hot topic today. Strengthening the construction of medical disciplines could not only promote the progress of medical technology, but also promote the development of interdisciplinary fields and provide reference for accelerating the Cross-Discipline of other disciplines. In order to intuitively show the similarity of disciplines, we use Gephi to draw a network map shown in Figure 4. Different colors indicate different communities. We could see that not only MATERIALS SCIENCE MULTIDISCIPLINARY and PHYSICS APPLIED have large Cross-Discipline phenomenon, but also ENGINEERING ELECTRICAL & ELECTRINIC, PSYCHOLOGY have large similarity with other disciplines. There is a complex Cross-Discipline between disciplines. We divide this network into several parts to analysis. Due to the space limitation, we just analyze the ENGINEERING ELECTRICAL & ELECTRINIC discipline. The size of the node indicates the status of the discipline in the network and the thickness of the connection indicates the degree of similarity, the greater the degree of similarity, the thicker the connection will be. We could intuitively observe that ENGINEERING ELECTRICAL ELECTRONIC, COMPUTER SCIENCE INFORMATION SYSTEMS, COMPUTER SCIENCE THEROY & METHODS and other disciplines have obvious connections between each other.
Figure 4. Cross-Discipline network in 2015. (The threshold is set to 0.2)

Figure 5. A part of Cross-Discipline network in 2015 (The threshold is set to 0.2)
Through the Figure 5, the size of the node indicates the status of the discipline in the network and the thickness of the connection indicates the degree of similarity, the greater the degree of similarity, the thicker the connection will be. We could intuitively observe that ENGINEERING ELECTRICAL ELECTRONIC, COMPUTER SCIENCE INFORMATION SYSTEMS, COMPUTER SCIENCE THEROY & METHODS and other disciplines have obvious connections between each other.

**Discipline overall vector**

We used all of the journals included in 234 disciplines as the overall discipline, and calculate the similarity between the overall discipline and other disciplines. Similarly, we use the Vector Space Model to calculate the degree of Cross-Discipline, the discipline overall vectors are shown in Table 3.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Disciplines</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BIOCHEMISTRY &amp; MOLECULAR BIOLOGY</td>
<td>0.27</td>
</tr>
<tr>
<td>2</td>
<td>ECONOMICS</td>
<td>0.27</td>
</tr>
<tr>
<td>3</td>
<td>MATERIALS SCIENCE MULTIDISCIPLINARY</td>
<td>0.26</td>
</tr>
<tr>
<td>4</td>
<td>MATHEMATICS APPLIED</td>
<td>0.26</td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>0.25</td>
</tr>
<tr>
<td>6</td>
<td>NEUROSCIENCES</td>
<td>0.25</td>
</tr>
<tr>
<td>7</td>
<td>ENGINEERING ELECTRICAL &amp; ELECTRONIC</td>
<td>0.24</td>
</tr>
<tr>
<td>8</td>
<td>PHARMACOLOGY &amp; PHARMACY</td>
<td>0.24</td>
</tr>
<tr>
<td>9</td>
<td>PHYSICS APPLIED</td>
<td>0.24</td>
</tr>
<tr>
<td>10</td>
<td>NANOSENCE &amp; NANO-TECHNOLOGY</td>
<td>0.23</td>
</tr>
<tr>
<td>11</td>
<td>CLINICAL NEUROLOGY</td>
<td>0.22</td>
</tr>
<tr>
<td>12</td>
<td>ENVIRONMENTAL SCIENCES</td>
<td>0.22</td>
</tr>
<tr>
<td>13</td>
<td>CELL BIOLOGY</td>
<td>0.21</td>
</tr>
<tr>
<td>14</td>
<td>BIOTECHNOLOGY &amp; APPLIED MICROBIOLOGY</td>
<td>0.21</td>
</tr>
<tr>
<td>15</td>
<td>PSYCHIATRY</td>
<td>0.21</td>
</tr>
<tr>
<td>16</td>
<td>GENETICS &amp; HEREDITY</td>
<td>0.20</td>
</tr>
<tr>
<td>17</td>
<td>ONCOLOGY</td>
<td>0.20</td>
</tr>
<tr>
<td>18</td>
<td>PUBLIC ENVIRONMENTAL &amp; OCCUPATIONAL HEALTH</td>
<td>0.20</td>
</tr>
<tr>
<td>19</td>
<td>CHEMISTRY PHYSICAL</td>
<td>0.19</td>
</tr>
<tr>
<td>20</td>
<td>MATHEMATICS INTERDISCIPLINARY APPLICATIONS</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Among the 234 disciplines of Web of Science, the discipline which has the largest similarity with all other disciplines is BIOCHEMISTRY & MOLECULAR BIOLOGY and the values of similarity are 0.27. The top 10 disciplines ranked by similarity mostly belong to nature science indicating that nature science played a basic role in supporting other discipline development. In addition, there are totally 89 disciplines whose discipline overall vector similarity are less than 0.1, accounting for 38.0% of all disciplines. And there are totally 145 disciplines whose discipline overall vector similarity are more than 0.1, accounting for 62.0% of all disciplines. It could be concluded that most disciplines are closely related to the development of the whole disciplines. The improvement of the whole disciplines not only improves the progress of the whole discipline, but also improves the development of each discipline.

**Conclusions**

With the rapid development of science, the phenomenon of independent study has gradually disappeared and more and more disciplines begin to converge. Different disciplines through Cross-Discipline with other disciplines could promote the development of the discipline itself and also promote the development of other disciplines. There is a trend that Cross-Discipline
becomes more and more common and disciplines are merging. Interdisciplinary often lead to the birth of new scientific achievements and subsequently accelerate the development of disciplines.

In order to measure the similarity between disciplines, we propose a Cross-Discipline Vector Space Model method based on the 234 disciplines of Web of Science. We find that a large degree of Cross-Discipline exists in different disciplines and the degree of similarity between discipline groups also reaches a high level. Through the research of the high degree of Cross-Discipline between disciplines, the high degree of Cross-Discipline between disciplines and discipline groups, and the high degree of Cross-Discipline between disciplines and the entire set of disciplines, we can conclude that disciplines have accomplish a large degree of Cross-Discipline merging.

We used the Vector Space Model to calculate the similarity between disciplines, and used the degree of Cross-Discipline values to determine the correlations. The larger the value is, the closer the connection is. The advantage of the quantitative evaluation method is more scientific to judge the development status of Cross-Discipline between disciplines, and could provide scientific suggestions for determine the trend of disciplines development and disciplinary activities in the future.

Disciplinary barriers are the main causes of impeding the integration of disciplines. Therefore, colleges and institutes have to break through the barriers when they want to cultivate comprehensive talents. Breaking through the existing talent training model will perfect the program of cultivating. The integration of these disciplines will provide some referential significance for promoting the level of scientific research and technology into the leader group in the world. For example: we could divide MATERIALS SCIENCE MULTIDISCIPLINARY, PHYSICS APPLIED and NANOSCIENCE & NANO TECHNOLOGY into a large discipline for promoting the development of MATERIALS SCIENCE MULTIDISCIPLINARY, which will be more conductive to the development of the whole discipline. At the same time, we should pay more attention to the Cross-Discipline between different disciplines and can not only confine ourselves to a specific field, and it will be easy to break through discipline barriers and achieve success. The development of Cross-Discipline is an inevitable trend. Strengthening the development of universities and research institutions and promoting disciplines merging will provide a referential significance for constructing a list of world-class universities and improving the speed of scientific research.

In addition, the trend of Cross-Discipline is still in the stage of continuous development, Disciplines in different fields are merging and interactive at all time. We attempt to measure the Cross-Discipline phenomenon in the top journals and magazines to find out the development direction of all scientific fields, and predict the development trend and opportunity in the future. It is important for researchers to know how to grasp the scientific development trend and guide us towards a more effective scientific research. We attempt to measure the degree of Cross-Discipline and transform the relevance of disciplines into information for researchers, in order to bring more scientific research values.

Acknowledgments
This work is partially supported by grant from the Natural Science Foundation of China (No. 61402075, 61572102, 61602078), the Ministry of Education Humanities and Social Science Project (No. 16YJCZH12), the Fundamental Research Funds for the Central Universities.

References


Consistency of interdisciplinarity indicators

Qi Wang\textsuperscript{1} Jesper Wiborg Schneider\textsuperscript{2}

\textsuperscript{1}qiwang@ps.au.dk \textsuperscript{2}jws@ps.au.dk
\textsuperscript{1,2}Aarhus University (Denmark)

Abstract
Assessing the interdisciplinarity of a given body of research is an important work in bibliometric studies. Since the nature and concept of interdisciplinary is ambiguous and uncertain, various indicators have been created to capture interdisciplinarity. However, few studies have examined the consistency of these indicators. In this context, this paper aims to systematically explore and compare to what extent these interdisciplinarity measures are consistent, and observe which attributes of interdisciplinarity they aim to indicate. We examine these interdisciplinarity measures on WoS categories. Based on the Pearson’s correlation coefficient, we find most measures in the different conceptual groups do not show high correlations. However, although measures in the same group are expected to be strongly related, some measures that are supposed to indicate the same attribute of interdisciplinarity are rather dissimilar. Furthermore, distribution figures provide an insight of the discriminatory power of the measures. Hence, we can identify measures with poor discriminatory power. In addition, we selected five WoS categories to carry out an in-depth analysis. These analyses are, to some extent, consistent with the Pearson’s correlation coefficients. Our results provide some evidence that these measures indicate different attributes of interdisciplinarity, perhaps more than presumed.

Conference Topic
Indicators, Science policy and research assessment, Citation and co-citation analysis

Introduction
Works that aim to quantitatively assess interdisciplinarity tend to bemoan the measurement situation, for there are no consensus on the definition and operationalization of interdisciplinary research. Since the concept of interdisciplinarity is ambiguous and not well delineated, yet several indicators purport to measure it or aspects of it, it is important to empirical examine the potential consistency of such measures in order to be able to more thoroughly analyse the meaning and validity of these constructs. However, while numerous indicators have been created to measure interdisciplinarity, empirical comparisons of interdisciplinarity measures are generally overlooked. In this context, this paper aims to systematically explore and compare to what extent that these interdisciplinarity measures are consistent, and observe which attributes of interdisciplinarity they aim to indicate.

Previous explorations of ways to measure interdisciplinarity can be carried out from various aspects, for instance analyzing publications, researchers, and terminologies. While we constrain our review to indicators based on bibliometric data such as publications, references, and citations, there remains a variety of other approaches to measuring interdisciplinarity. This study is organized as follows. We first provide some related work on the conceptual and analytical frameworks of interdisciplinarity. Then, we summary previous measures of interdisciplinarity. Next section briefly introduces the data in this work. The following section reports the results of our analysis. Discussion and conclusions is in the last section.

Analytical framework of interdisciplinarity
Regarding the conceptual and analytical framework of interdisciplinarity measures, Rafols and Meyer (2010) develop the conceptual framework of interdisciplinarity by exploring the concepts of “diversity” and “coherence” in a bibliometric context. Diversity aims to “describe the heterogeneity of a bibliometric set” (Rafols & Meyer, 2010, p.263), and coherence include
indices that “are constructed to measure the intensity of similarity relations within a bibliometric set” (Rafols & Meyer, 2010, p.263). The assumption is that a cluster of publications with a high degree of interdisciplinarity, should contain a considerable number of within cluster citations spanning distant predefined journal subject categories. Compared to multidisciplinary, interdisciplinarity is seen as a more advanced stage of the relationship between disciplines in which integration of theories, methods, techniques and data is attained. As illustrated in Figure 1, interdisciplinarity is associated with high degree of diversity and coherence.

![Figure 1. Conceptual framework of IDR (Source: Rafols et al., 2012)](image)

In addition, Rafols et al. (2012) supplements the analyses of interdisciplinarity with a further concept of “intermediation”. Intermediation depicts a situation where a large proportion of publications in a set is in an “intermediate” position as depicted in Figure 1. This is another perspective of exploring interdisciplinarity, which stresses the position of a set of publications in a citation network. Whereas, the combination of diversity and coherence is more focused on the internal distinct components and their relations within a given publication set. As mentioned by Rafols and Meyer (2010), their conceptual framework is constructed from the perspective of a knowledge integration process.

**Overview of interdisciplinarity measures**

This section introduces 16 bibliometric-based measures typically operationalised as one-dimensional (non-composite) indicators of interdisciplinarity. In order to present these measures systematically, they are embedded into Rafols and colleagues’ conceptual framework of interdisciplinarity. Another point should be elucidated before introducing these measures, the issue of unit of analyses. This work examines the interdisciplinarity of WoS categories, which we will introduce in the next section. Outgoing (references) and incoming citations (citations) can both be used to analyse the interdisciplinarity of a publication set. However, “there does not seem to be clear evidence that one is preferable to the other” (Levitt et al., 2011, p.1121), while others stress that citations and references have different implications (e.g. Porter & Chubin, 1985). To simplify, we limit this study to the references in publications, i.e. outgoing citations. Besides this, it should be noted that not all bibliometric-based indicators are included in the present work. This is because not all indicators are suitable for measuring interdisciplinarity of WoS categories.

---

1 Notice, these measures were all developed using the Web of Science database.
Diversity

As mentioned above, diversity includes measures that “describe the heterogeneity of a bibliometric set” (Rafols & Meyer, 2010, p.263). We first introduce measures depending on the WoS journal classification system. WoS allows journals to have multiple categories, and thus researchers can use this characteristic to assess interdisciplinarity. Morillo et al. (2001, 2003) propose several measures to describe the interdisciplinarity based on the WoS classification system. Table 1 summarizes some of their measures.

It is obvious that the arbitrary construction of the WoS classification system seriously influences the measurement of diversity. Therefore, some researchers are focused on measuring the proportion of references to outside categories (PRO). It presumes that the higher proportion of references to journals in other categories, the larger the expectation that the journal contains publication with interdisciplinarity research. Nevertheless, PRO is still a rather simple indicator that lacks considerations on the distribution of references over categories, the number of outside categories, similarity between each pair of categories, etc. Thus, relatively complex measures have been continuously proposed to improve the measurement of diversity. For instance, some researchers prefer to adopt a probability perspective, such as Simpson’s diversity (Simpson, 1949), Shannon’s entropy (e.g. Stirling, 2007; Leydesdorff & Rafols, 2011; Silve et al., 2013) and Brillouin’s index (Steele & Stier, 2000, see also Brillouin, 1956; Chang & Huang, 2012). Economic measures have also been used as indicators of interdisciplinarity. For instance, the Gini coefficient, which is perceived as a measure of income inequality, is used to evaluate the degree of unevenness in citation or reference distributions of subject categories (i.e. Leydesdorff & Rafols, 2011). Besides Gini coefficients, other indicators have also been adopted and proposed to measure the concentration of publications in a research field, for instance, Pratt index (Pratt, 1977, see also Morillo et al., 2001) and Specialization measurement (Porter et al., 2007; 2008), as shown in Table 1. Note that these three concentration-related measures are inversely related with interdisciplinarity.

The diversity measures introduced so far focus on the distribution of citations over different categories, but they are criticized for not considering the similarity of pairs of such categories. Rao (1982) and Stirling (2007) respectively proposed that similarity should be taken into account when assessing diversity. Here we need to introduce some mathematical notions here. Let $c_{ik}$ denote the number of references from WoS subject category $i$ to $k$, and $t_i$ denote the total number of reference for category $i$, then we have $t_i = \sum_{k=1}^{n} c_{ik}$. The percentage of references from category $i$ to $k$ can be expressed as $p_{ik} = \frac{c_{ik}}{t_i}$. Then the general diversity formula is as follows,

$$\Delta = \sum_{i \neq j} (d_{ij})^\alpha (p_{ik}p_{jk})^\beta,$$

in which, $d_{ij}$ shows the difference between the category $i$ and $j$. Other traditional indices can be derived by setting the value of parameters $\alpha$ and $\beta$. For example, when we define delta “as the variant for $\alpha = 1, \beta = 1$, [we get] the simplest form incorporating variety, balance and similarity” (Rafols & Meyer, 2010, p. 267). This is known as the Rao-Stirling index:

$$\text{Rao – Stirling index} = \sum_{i \neq j} d_{ij}p_{ij}p_{ij}.$$

The Rao-Stirling measure has been widely used to measure diversity and thus been used as an indicator for interdisciplinarity (e.g. Porter et al., 2007; Porter & Rafols, 2009; Wang et al., 2015). Note that there are several ways to construct the matrix of dissimilarity. While Salton’s cosine similarity (Salton & McGill, 1983) is frequently applied, it has several transformations. Hence, different results can be expected (e.g., Schneider & Borlund, 2007a; 2007b). Here, we discuss two transformations of the cosine formula to calculate similarity. Suppose we aim to attain a symmetric similarity matrix of WoS categories $[s_{ij}]$ by using their citation relations. The first step is to construct a matrix of citation relations between categories $[c_{ij}]$. Note that
[s_{ij}] is a symmetric matrix, but [c_{ij}] is asymmetric. We use \( c_{ij} \) to denote the number of citations from publications in \( i \) to publications in \( j \).

The first way of using Salton’s cosine index can be illustrated as two subject categories are considered to be strongly related if they commonly cite the same subject categories. Hence, the similarity of two subject categories \( i \) and \( j \) is given by

\[
S_c(i,j) = \frac{\Sigma_k c_{ik} c_{jk}}{\sqrt{(\Sigma_k c_{ik}^2)(\Sigma_k c_{jk}^2)}}
\]

Another way of using the cosine formula (also called the Ochiai index) is that

\[
S_O(i,j) = \frac{c_{ij} + c_{ji}}{\sqrt{(\Sigma_k c_{ik} + \Sigma_k c_{ki})(\Sigma_k c_{kj} + \Sigma_k c{kj})}}, \quad i \neq j
\]

Here, \( c_{ij} + c_{ji} \) is equal to the total number of citations between subject category \( i \) and \( j \). Note that \( S_O(i,i) \) is set as 1. (More details, see Zhang et al., 2016).

In addition, several strategies can transform a similarity matrix into a dissimilarity matrix. The mainstream solution is to use \( 1 - [s_{ij}] \) to attain a distance matrix \([d_{ij}]\). There are also studies using \( 1/[s_{ij}] \) to construct a distance matrix (e.g. Jensen & Lutkouskaya, 2014). This work will examine both strategies of constructing a dissimilarity matrix.

Besides this, it should be noted that Rao-Stirling diversity can be used at various levels. To be more specific, one could measure Rao-Stirling diversity for each publication in a category and then calculate the mean value as the interdisciplinarity of this category. On the other hand, one could also view this category as a whole, calculating the proportion of references in all publications over different categories. In this work, we examine Rao-Stirling index at both levels. Table 1 summarizes different transformations of the Rao-Stirling index, where rs_p is based on the references of individual publications, and rs_r is based on the references of all publications in a dataset.

Recently, Zhang et al. (2016) claim that the Rao-Stirling measure has a low discriminatory power, since the interdisciplinarity results obtained by Rao-Stirling measure are quite similar. Instead, they promote the Hill type measure \( ^2D^2 \) to overcome these limitations. The formula of \( ^2D^2 \) is shown in Table 1.

### Coherence

We now move to another conceptual group of measures, namely coherence, which is used to capture the intensity of a group of publications. Note that, according to the studies by Rafols and colleagues (Rafols & Meyer, 2010; Rafols, 2014), diversity and coherence should be used in conjunction to assess interdisciplinarity. However, in the light of our summary of previous studies, coherence tends to be ignored when measuring interdisciplinarity. Many studies use merely diversity to indicate interdisciplinarity (e.g. Porter & Rafols, 2009). The reason might be that coherence has often been conceived as an inherent attribute within the dataset under study. For instance, if research topics are identified using a certain clustering algorithm, they would be expected to be coherent.

As mentioned, the present study will measure the interdisciplinarity of WoS categories. In the present study, we make a strong assumption that publications in the same WoS category are expected to display strong coherence\(^2\). Hence, we will only test one coherence measure, which is shown in Table 1. However, it is worth noticing that some studies claim that this coherence measure is suitable for quantifying the degree interdisciplinarity (more details, see Soo & Kampis, 2012; Wang, 2016).

\(^2\) This is most likely not true, for all publications or categories, given the well-known core and scatter phenomena in literatures.
Intermediation

As discussed above, Rafols et al. (2012) supplement their analyses of interdisciplinarity with the concept of “intermediation”. However, measures of intermediation designed for interdisciplinary purposes are few. One frequently used index is the “betweenness centrality” measure, which is originally proposed by Freeman (1977) and later introduced by Leydesdorff (2007) to measure the interdisciplinarity of journals. Two other measures of intermediation are introduced by Rafols et al. (2012). The first is a clustering coefficient, which “identifies the proportion of observed links between journals over the possible maximum number of links” (de Nooy et al., 2005, p.149; see also Rafols et al., 2012, p. 1268). The second measure of intermediation proposed by Rafols and colleagues is “the average similarity of a given journals to all other journals” (Rafols et al., 2012, p. 1268).

### Table 1. Measures of interdisciplinarity

<table>
<thead>
<tr>
<th>ID</th>
<th>Measure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>p_multi</td>
<td>The percentage of multi-assigned journals in a category.</td>
</tr>
<tr>
<td>2</td>
<td>p_in_so</td>
<td>The percentage of journals that are classified in another category inside the area. Research areas are a higher aggregation level of WoS categories.</td>
</tr>
<tr>
<td>3</td>
<td>p_out_so</td>
<td>The percentage of journals that are classified in another category outside the area.</td>
</tr>
<tr>
<td>4</td>
<td>d_links</td>
<td>The number of links between different categories established by journals in a given category</td>
</tr>
<tr>
<td>5</td>
<td>PRO</td>
<td>$c_{ik}/\sum_{k=1 \neq k}^{n} c_{ik}$, in which $k$ is the index attained by sorting subject categories according to $c_{ik}$ in non-decreasing order.</td>
</tr>
<tr>
<td>6</td>
<td>Simpson’s index</td>
<td>$1 - \sum_{k=1}^{n} p_{ik}^2$</td>
</tr>
<tr>
<td>7</td>
<td>Shannon’s entropy</td>
<td>$-\sum_{k=1}^{n} p_{ik} \log p_{ik}$</td>
</tr>
<tr>
<td>8</td>
<td>Brillouin’ index</td>
<td>$(\log t_i! - \sum (\log c_{ik}!))/t_i$</td>
</tr>
<tr>
<td>9</td>
<td>Gini coefficient</td>
<td>$\sum_{k=1}^{n} (2k-n-1)c_{ik}/n \sum_{k=1}^{n} c_{ik}$, in which $k$ is the sequence number obtained by ranking $p_{ik}$ in decreasing order.</td>
</tr>
<tr>
<td>10</td>
<td>Pratt index</td>
<td>$2^{(n+1)/2} q^{1/2}$, where $q = \sum_{k=1}^{n} kp_{ik}$ and $k$ is the sequence number obtained by ranking $p_{ik}$ in decreasing order.</td>
</tr>
<tr>
<td>11</td>
<td>Specialization</td>
<td>$\sum_{k=1}^{n} c_{ik}^2 / (\sum_{k=1}^{n} c_{ik})^2$</td>
</tr>
<tr>
<td>12</td>
<td>rs_p [1-Sc]</td>
<td>$(1/n) \sum_{i \neq j} (1 - S_{c(s_i,s_j)}) p_i p_j$</td>
</tr>
<tr>
<td>13</td>
<td>rs_r [1-Sc]</td>
<td>$\sum_{i \neq j} (1 - S_{c(s_i,s_j)}) p_i p_j$</td>
</tr>
<tr>
<td>14</td>
<td>rs_p [1/Sc]</td>
<td>$(1/n) \sum_{i \neq j} (1/S_{c(s_i,s_j)}) p_i p_j$</td>
</tr>
<tr>
<td>15</td>
<td>rs_r [1/Sc]</td>
<td>$\sum_{i \neq j} (1/S_{c(s_i,s_j)}) p_i p_j$</td>
</tr>
<tr>
<td>16</td>
<td>rs_p [1-So]</td>
<td>$(1/n) \sum_{i \neq j} (1 - S_{o(s_i,s_j)}) p_i p_j$</td>
</tr>
<tr>
<td>17</td>
<td>rs_r [1-So]</td>
<td>$\sum_{i \neq j} (1 - S_{o(s_i,s_j)}) p_i p_j$</td>
</tr>
<tr>
<td>18</td>
<td>Hill Type [Sc]</td>
<td>$1/ \sum_{i \neq j} S_{c(s_i,s_j)} p_i p_j$</td>
</tr>
<tr>
<td>19</td>
<td>Coherence</td>
<td>$\sum_{i \neq j} p_i d_{ij}$, where $p_{ij}$ is the proportion of references from category $i$ to $j$</td>
</tr>
<tr>
<td>20</td>
<td>Betweenness centrality (BC)</td>
<td>$\sum_i \sum_{j \neq k} c_{ikj}/c_{ij}$, in which $c_{ikj}$ is the number of references from category $i$ to $j$ pass through category $k$.</td>
</tr>
<tr>
<td>21</td>
<td>Cluster coefficient (CC)</td>
<td>$\sum_i p_i \frac{c_{ij}}{a_i a_j}$, in which $a_i$ is the number of publications that category $i$ has.</td>
</tr>
</tbody>
</table>
Average similarity (AS) \[ \sum_i p_i \left( \frac{1}{N} \sum_j s_{ij} \right), \]

Data

This paper aims to systematically explore and compare to what extent these interdisciplinarity measures are consistent, and further observe which attributes of interdisciplinarity they aim to indicate. As a case, we examine the interdisciplinarity of WoS subject categories using the measures that we have introduced in the previous section. As data we use a publication set consisting of all publications of the document type article published in 2010 from the in-house WoS database. Our study uses publications and their references to investigate interdisciplinarity. The validity and effectiveness of using this type of data source might be questioned in some research fields, especially in the fields where journals are not the main scientific communication medium. We therefore exclude journals from the Arts & Humanities Citation Index resulting in a total of 224 WoS categories included in the analysis.

One issue worth highlighting. The WoS classification system is a coarse classification of science; however, some measures might be problematic when working at the high aggregation levels of science. For instance, measures of intermediation like “betweenness centrality” are sensitive to the level of aggregation, and it should probably be applied at levels lower than WoS categories (Rafols et al., 2012). Besides this, it would be more reasonable to measure coherence when evaluating the interdisciplinarity of an institution. Since publications of an academic institution might be diverse, but they do not closely connect with each other.

Hence, examining these measures at various levels of aggregation seems necessary. However, due to space constrains, the present work will only examine the measures that we introduced in the previous section and are relevant to measuring the interdisciplinarity of WoS categories. Future studies will expand these analyses to different units of analysis and at different levels.

Results

This section presents the results of our analysis. We will first provide the results regarding the relations of the interdisciplinarity measures, then report the distribution over categories of each measure, and finally provide an in-depth analysis with five categories.

Relations between indicators of interdisciplinarity

Pearson’s correlation is used to analyse the linear correlations between the measures of interdisciplinarity. From the previous discussion, we know that these measures can be roughly classified into six groups according to the attribute of interdisciplinarity they aim to quantify, see Tables 1. Table 2 provides Pearson’s correlation coefficients, and the results for each group is framed in black.

First, it can be seen that the measures in the first group to some extent are highly correlated with each other, except p_in_so. As mentioned, p_in_so is defined as the percentage of journals that are classified in another category inside a research area. In this paper, we adopt the strategy provided by CWTS, which assigns WoS categories into 35 research areas. Hence, p_in_so seems to constrain the diversity of journals of a category into a given research field. Thus, it is not surprising for it to have a weak correlation with other measures in its group and even hold a negative correlation with p_out_so. The first four measures in the second group, i.e. entropy, Brillouin, Simpson and PRO, are highly correlated, especially Shannon entropy and Brillouin index. This is in line with our expectation, since Brillouin’s index is mathematically related to Shannon’s entropy (Steele & Stier, 2000). The third group is designed to capture the degree of centralized tendency of publications in the same category. Therefore, we expect them to have a negative or weak relation with other measures of interdisciplinarity. However, Pratt index
differs from our expectation, for it shows a negative relation with the Gini coefficient and Specialization. Pratt index seems to quantify the dispersion degree of publications in a category, and should probably therefore be assigned to the first group.

Measures in the fourth group complement previous measures by considering distance. As shown in Table 1, using different similarity measures, namely $S_c$ and $S_o$, lead to comparatively weakly related Rao-Stirling results. On the other hand, and not surprisingly, Rao-Stirling measures using the same similarity measure at the level of individual publications and categories are strongly linearly associated with each other. In short, $rs_p$ and $rs_r$ with the same similarity matrix are highly correlated. However, the comparative correlation profile of the Hill type measure differs with all other Rao-Stirling measures. While it applies the same cosine function as $rs_p$ [1-$S_o$] and $rs_r$ [1-$S_o$], their correlations are weak.

Coherence has relatively weak correlations with all other measures, except the Rao-Stirling [$S_c$] version. This seems reasonable as the coherence measure is conceptually different, and strong coherence may not imply a high degree of interdisciplinarity. Finally, the left three measures focus on intermediation. Intermediation differs from coherence, since a high intermediation of a WoS category may suggest that the category has a high degree of interdisciplinarity, as shown in Figure 1. Our results show that these three measures are weakly correlated with all other measures.

In addition, we also examine to what extent measures in the different groups quantify different attributes of interdisciplinarity. To provide more insight to this issue, a cluster solution is produced using a Ward’s cluster algorithm. As shown in the dendrogram in Figure 2, measures below the line are perceived to describe the diversity attribute, whereas the others mainly include measures regarding intermediation; i.e. the latter reflects the proportion of publications of a category in “intermediate positions”, and the centralized tendency of publications within a category.

Our results indicate that these measures, so far in the literature, have been operationalized (and perhaps interpreted) rather differently when used as indicators of interdisciplinarity. On the other hand, some measures are clustered into the group that seems unintuitive. For instance, the Hill-type index aims to capture the attribute of diversity, but it presents a significant difference with other diversity measures.

Table 2. Correlation table of 16/22 measures used as indicators of interdisciplinarity

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{multi}$</td>
<td>1.00</td>
<td>0.79</td>
<td>0.41</td>
<td>0.63</td>
<td>0.19</td>
<td>0.22</td>
<td>0.29</td>
<td>0.44</td>
<td>-0.36</td>
<td>0.23</td>
<td>-0.83</td>
</tr>
<tr>
<td>$p_{out_so}$</td>
<td>0.79</td>
<td>1.00</td>
<td>-0.24</td>
<td>0.65</td>
<td>0.37</td>
<td>0.39</td>
<td>0.40</td>
<td>0.45</td>
<td>-0.30</td>
<td>0.02</td>
<td>-0.72</td>
</tr>
<tr>
<td>$p_{in_so}$</td>
<td>0.41</td>
<td>-0.24</td>
<td>1.00</td>
<td>0.02</td>
<td>-0.25</td>
<td>-0.24</td>
<td>-0.15</td>
<td>0.03</td>
<td>-0.13</td>
<td>0.35</td>
<td>-0.25</td>
</tr>
<tr>
<td>$d_{links}$</td>
<td>0.63</td>
<td>0.65</td>
<td>0.02</td>
<td>1.00</td>
<td>0.43</td>
<td>0.44</td>
<td>0.47</td>
<td>0.56</td>
<td>-0.21</td>
<td>-0.14</td>
<td>-0.55</td>
</tr>
<tr>
<td>entropy</td>
<td>0.19</td>
<td>0.37</td>
<td>-0.25</td>
<td>0.43</td>
<td>1.00</td>
<td>1.00</td>
<td>0.86</td>
<td>0.64</td>
<td>-0.26</td>
<td>-0.55</td>
<td>-0.31</td>
</tr>
<tr>
<td>Brillouin</td>
<td>0.22</td>
<td>0.39</td>
<td>-0.24</td>
<td>0.44</td>
<td>1.00</td>
<td>1.00</td>
<td>0.86</td>
<td>0.64</td>
<td>-0.31</td>
<td>-0.49</td>
<td>-0.35</td>
</tr>
<tr>
<td>Simpson</td>
<td>0.29</td>
<td>0.40</td>
<td>-0.15</td>
<td>0.47</td>
<td>0.86</td>
<td>0.86</td>
<td>1.00</td>
<td>0.83</td>
<td>-0.12</td>
<td>-0.42</td>
<td>-0.39</td>
</tr>
<tr>
<td>PRO</td>
<td>0.44</td>
<td>0.45</td>
<td>0.03</td>
<td>0.56</td>
<td>0.64</td>
<td>0.64</td>
<td>0.83</td>
<td>1.00</td>
<td>-0.05</td>
<td>-0.33</td>
<td>-0.39</td>
</tr>
<tr>
<td>Gini</td>
<td>-0.36</td>
<td>-0.30</td>
<td>-0.13</td>
<td>-0.21</td>
<td>-0.26</td>
<td>-0.31</td>
<td>-0.12</td>
<td>-0.05</td>
<td>1.00</td>
<td>-0.56</td>
<td>0.53</td>
</tr>
<tr>
<td>Pratt</td>
<td>0.23</td>
<td>0.02</td>
<td>0.35</td>
<td>-0.14</td>
<td>-0.55</td>
<td>-0.49</td>
<td>-0.42</td>
<td>-0.33</td>
<td>-0.56</td>
<td>1.00</td>
<td>-0.29</td>
</tr>
<tr>
<td>Spec</td>
<td>-0.83</td>
<td>-0.72</td>
<td>-0.25</td>
<td>-0.55</td>
<td>-0.31</td>
<td>-0.35</td>
<td>-0.39</td>
<td>-0.39</td>
<td>0.53</td>
<td>-0.29</td>
<td>1.00</td>
</tr>
<tr>
<td>$rs_p$ [1-$S_c$]</td>
<td>0.35</td>
<td>0.49</td>
<td>-0.17</td>
<td>0.38</td>
<td>0.37</td>
<td>0.36</td>
<td>0.37</td>
<td>0.36</td>
<td>0.03</td>
<td>-0.32</td>
<td>-0.27</td>
</tr>
<tr>
<td>$rs_r$ [1-$S_c$]</td>
<td>0.22</td>
<td>0.40</td>
<td>-0.25</td>
<td>0.33</td>
<td>0.50</td>
<td>0.48</td>
<td>0.40</td>
<td>0.32</td>
<td>0.06</td>
<td>-0.50</td>
<td>-0.18</td>
</tr>
<tr>
<td>$rs_p$ [1/$S_c$]</td>
<td>0.25</td>
<td>0.34</td>
<td>-0.10</td>
<td>0.23</td>
<td>0.16</td>
<td>0.14</td>
<td>0.08</td>
<td>0.16</td>
<td>0.06</td>
<td>-0.24</td>
<td>-0.16</td>
</tr>
<tr>
<td>$rs_r$ [1/$S_c$]</td>
<td>0.26</td>
<td>0.37</td>
<td>-0.14</td>
<td>0.27</td>
<td>0.30</td>
<td>0.29</td>
<td>0.17</td>
<td>0.19</td>
<td>0.02</td>
<td>-0.31</td>
<td>-0.18</td>
</tr>
<tr>
<td>$rs_p$ [1-$S_o$]</td>
<td>0.00</td>
<td>0.18</td>
<td>-0.26</td>
<td>0.28</td>
<td>0.60</td>
<td>0.59</td>
<td>0.60</td>
<td>0.40</td>
<td>-0.01</td>
<td>-0.36</td>
<td>-0.06</td>
</tr>
<tr>
<td>$rs_r$ [1-$S_o$]</td>
<td>0.03</td>
<td>0.22</td>
<td>-0.28</td>
<td>0.29</td>
<td>0.66</td>
<td>0.65</td>
<td>0.61</td>
<td>0.39</td>
<td>-0.01</td>
<td>-0.43</td>
<td>-0.07</td>
</tr>
</tbody>
</table>

1412
Discriminatory power is one criteria for evaluating the effectiveness of indicators. As mentioned in Rafols and Meyer (2010), “since we do not have benchmarks for diversity or coherence from other areas of science, we cannot investigate whether this field is particularly
diverse or coherent” (p. 275). In other words, it is challenging to quantitatively determine if an indicator of interdisciplinarity is actually “valid”. However, distribution figures can implicitly provide the impression of degrees of interdisciplinarity in the entire science system. In other words, we can estimate the number of WoS categories with presumably high or low degrees of interdisciplinarity according to measure applied, and then heuristically assess if the results are in line with our expectation. Distributions of WoS categories over “interdisciplinarity” are provided in Figure 3. For each graph, x-axis shows the degree of “interdisciplinarity” and y-axis shows the number of categories.

As shown, the interdisciplinarity values of Simpson, Gini, Rao-Stirling index with $S_o$, Hill type measure, clustering coefficient, and average similarity are comparatively concentrated in a certain range. Hence, these indicators might be problematic in their discriminatory power. Taking the average similarity measure as an example, while it shows a normal distribution in Figure 3, its values actually range between zero and 0.3. Thus, we conclude this measure shows an unsatisfied discriminatory power. The remaining measures to some extent follow Gaussian distribution. Nevertheless, some are highly right-skewed like PRO and Pratt, and others are highly left-skewed like $p_{in \_so}$, $rs_p$ with $1/S_o$, and betweenness centrality.

![Figure 3. Distribution of interdisciplinarity indicators](image)

In-depth analysis for several WoS subject categories

A specific examination of the degree of interdisciplinarity of the WoS categories based on the various measures provide a more direct impression regarding the effectiveness of these measures. However, because of limited space, here we only examine five WoS categories in-depth; these are NANOSCIENCE & NANOTECHNOLOGY (NANO) and BIOCHEMISTRY & MOLECULAR BIOLOGY (BIOM), which previous studies often consider to be highly interdisciplinary, LAW and MATHEMATICS (MATH), which are supposed to show a low degree
of interdisciplinary, and INFORMATION SCIENCE & LIBRARY SCIENCE (LIS). LIS is chosen because readers of this paper might be familiar with this field.

Table 3 presents the interdisciplinarity ranks of these five subject categories obtained using the measures that we have introduced previously. Instead of reporting the interdisciplinarity values obtained from the measures, we list the rank number attained after sorting all 224 categories according to their interdisciplinarity scores in a non-decreasing order. It should be mentioned that Gini, Pratt, and Speciation are inversely related with interdisciplinarity.

Obviously, the results obtained using these measures are quite different. As explained previous, measures in the first group is based on the WoS journal classification system. Hence, they reflect to what extent journals in a category are “diverse”, or can be assigned into multiple categories. The second group generally focuses upon references outside the journal subject categories or over various categories. Shannon entropy, Brillouin index and Simpson diversity show similar results, whereas PRO show some differences compared to the other three. The next group describes the centralized tendency of publications within a category. Interestingly, the results of the three measure are rather dissimilar. This is consistent with our Pearson’s correlation result, as shown in Table 2. This may suggest that our classification of these measures are deficient, or that they conceptually measures more diverse constructs than hitherto presumed. The fourth group contains several transformations of the Rao-Stirling index, and they are expected to be consistent in their empirical results. However, the results are not in line with the expectations. Using varying dissimilarity (or distance) matrices and working at different unit levels lead to rather divergent results. Finally, results from the intermediation measures are also quite varied. Meanwhile, note that the clustering coefficient and average similarity measures have poor discriminatory power of interdisciplinarity scores; hence, their ranks are not so useful in this context.

In short, these measures seemingly provide very different results when it comes to degree of interdisciplinarity for categories. This we consider strong evidence for the claim that interdisciplinarity cannot be sufficiently captured by one measure as a one-dimensional non-composite indicator.

Table 3. Interdisciplinarity ranks of the five categories

<table>
<thead>
<tr>
<th></th>
<th>NANO</th>
<th>BIOM</th>
<th>LIS</th>
<th>LAW</th>
<th>MATH</th>
</tr>
</thead>
<tbody>
<tr>
<td>p_multi</td>
<td>6</td>
<td>60</td>
<td>185</td>
<td>177</td>
<td>186</td>
</tr>
<tr>
<td>p_out_so</td>
<td>32</td>
<td>104</td>
<td>140</td>
<td>133</td>
<td>214</td>
</tr>
<tr>
<td>p_in_so</td>
<td>40</td>
<td>35</td>
<td>214</td>
<td>206</td>
<td>37</td>
</tr>
<tr>
<td>d_links</td>
<td>41</td>
<td>161</td>
<td>176</td>
<td>169</td>
<td>221</td>
</tr>
<tr>
<td>entropy</td>
<td>171</td>
<td>73</td>
<td>82</td>
<td>141</td>
<td>224</td>
</tr>
<tr>
<td>Brillouin</td>
<td>169</td>
<td>71</td>
<td>80</td>
<td>141</td>
<td>224</td>
</tr>
<tr>
<td>Simpson</td>
<td>132</td>
<td>103</td>
<td>104</td>
<td>199</td>
<td>223</td>
</tr>
<tr>
<td>PRO</td>
<td>22</td>
<td>170</td>
<td>143</td>
<td>216</td>
<td>223</td>
</tr>
<tr>
<td>Gini</td>
<td>150</td>
<td>208</td>
<td>187</td>
<td>203</td>
<td>125</td>
</tr>
<tr>
<td>Pratt</td>
<td>12</td>
<td>128</td>
<td>141</td>
<td>113</td>
<td>2</td>
</tr>
<tr>
<td>Spec</td>
<td>219</td>
<td>224</td>
<td>90</td>
<td>42</td>
<td>25</td>
</tr>
<tr>
<td>rs_p [1-Sc]</td>
<td>145</td>
<td>193</td>
<td>30</td>
<td>216</td>
<td>224</td>
</tr>
<tr>
<td>rs_r [1-Sc]</td>
<td>191</td>
<td>205</td>
<td>13</td>
<td>77</td>
<td>224</td>
</tr>
<tr>
<td>rs_p [1/Sc]</td>
<td>160</td>
<td>201</td>
<td>7</td>
<td>100</td>
<td>138</td>
</tr>
<tr>
<td>rs_r [1/Sc]</td>
<td>176</td>
<td>180</td>
<td>11</td>
<td>51</td>
<td>111</td>
</tr>
<tr>
<td>rs_p [1-So]</td>
<td>189</td>
<td>153</td>
<td>114</td>
<td>101</td>
<td>224</td>
</tr>
<tr>
<td>rs_r [1-So]</td>
<td>208</td>
<td>168</td>
<td>96</td>
<td>39</td>
<td>223</td>
</tr>
<tr>
<td>Hill [So]</td>
<td>215</td>
<td>174</td>
<td>35</td>
<td>10</td>
<td>135</td>
</tr>
<tr>
<td>Coh [Sc]</td>
<td>209</td>
<td>214</td>
<td>19</td>
<td>152</td>
<td>224</td>
</tr>
</tbody>
</table>
Discussion and Conclusions

This paper aims to systematically explore and compare to what extent 16 interdisciplinarity measures are consistent, and to examine which attributes of interdisciplinarity they aim to indicate. We examine these interdisciplinarity measures in relation WoS categories. Based on the Pearson’s correlation coefficient, we find that most measures in the various groups do not have strong correlations. However, although measures in the same groups are expected to be strongly related, some measures that supposedly indicate the same attribute of interdisciplinarity turn out to be rather dissimilar. Furthermore, distribution figures provide an insight into the discriminatory power of the measures. Hence, we can identify measures with poor discriminatory power, for instance, measures of cluster coefficients and average similarity. These measures may not appropriate to be used. However, this may also cause by the choice of our dataset. In addition, we selected five WoS categories to carry out an in-depth analysis of the consistency. These results are to some extent consistent with the findings based on the Pearson’s correlation coefficients. Our results provide evidence for the claim that these measures indicate different attributes of interdisciplinarity, they are much more diverse than usually perceived.

Conducting such a study is challenging. Several issues are worth further discussion. First, we embed these interdisciplinarity measures into the analytical framework of Rafols and colleagues, since the framework can help us better understand various perspectives of interdisciplinarity. However, it does imply that this framework is not problematic. It is still open to discussion whether interdisciplinarity should be described using one or several attributes and thus measures. Our work provide some evidence that different measures of interdisciplinarity are not consistent, and they may actually capture different dimensions of interdisciplinarity. However, the question of which attributes are essential for depicting the nature of interdisciplinarity is still not answered.

Second, measures purporting to indicate the same attribute should be consistent. More specifically, we discuss several ways of using the Rao-Stirling measure, and they are expected to be consistent. However, the results are not in line with such expectations. It should be noted that using different dissimilarity (or distance) matrices and working at different levels of units would lead to different scores. Unfortunately, this issue is always overlooked in previous studies.

The main limitation of this work is that we have only examined the degree of interdisciplinarity for WoS categories. It may be that this is not an appropriate level for studying interdisciplinary research. Besides this, measures regarding coherence and intermediation do not perform well when examining the interdisciplinarity of WoS categories. Hence, further studies will be carried out at the disaggregate level of for example research institutions. In addition, our grouping of the measures maybe not be sufficiently appropriate. For instance, Shannon entropy, Brillouin index, and Gini coefficient probably should be assigned to the same group, for they all quantify unevenness in a dataset. Further studies should re-think and improve this grouping.

References


Interdisciplinary Knowledge Flow: across Disciplines and Scientists in Scientific Funding

Jiang Wu¹, Ke Dong¹, Xiu-hao Ding²

¹ School of Information Management, Wuhan University, Wuhan, China, 430072
² School of Management, Huazhong University of Science and Technology, Wuhan, China, 430074
Email: jiangw@whu.edu.cn

Abstract

In scientific funding applications, scientists usually pursue interdisciplinary applications that lead to interdisciplinary knowledge flow. This paper utilizes the co-occurrences of disciplinary application codes (DACs) to construct an interdisciplinary knowledge flow network. Based on 193517 sponsored projects of the National Natural Science Foundation of China (NSFC), we study the interdisciplinary flow of knowledge. Results show that the interdisciplinary knowledge flow network is a small-world network. The distribution of knowledge flowability follows the 80/20 law. Two main knowledge flow paths across scientific departments exist, showing the heterogeneity of knowledge distributions across scientific disciplines. Furthermore, it is also found that individual disciplinary diversity in the interdisciplinary knowledge flow can significantly positively influences the number of sponsored projects and the total sponsored money. The findings can give implications to the policy interventions in research practices.

Conference Topic
Science policy (on different levels), Studies on the level of individual scientists, Co-occurrence analysis, Social network analysis

Introduction

In scientific communities, knowledge flows across various scientific disciplines to solve emerging complex problems (Siedlok, Hibbert, & Sillince, 2015). Science is indeed becoming more interdisciplinary (Porter & Rafols, 2009) and scientists themselves need interdisciplinary knowledge and seek a variety of knowledge from other individuals through interdisciplinary collaborations (Carayol & Thi, 2005). In scientific funding, scientists sometimes pursue interdisciplinary scientific funding applications to succeed in academic competitions. Generally, interdisciplinary scientists are more competitive in applications. They do interdisciplinary research and have abilities to integrate knowledge flexibly. The relationship between scientific disciplines is strongly influenced by national funding foundations (Lyall, Bruce, Marsden, & Meagher, 2013). Interdisciplinary research needs financial support from different disciplines (van Rijnsoever & Hessels, 2011), and many scientific foundations, including the National Natural Science Foundation of China (NSFC), encourage interdisciplinary collaborations to accelerate innovation (Benner & Sandström, 2000). Understanding the rules of interdisciplinary knowledge flow of scientific funding is important to help create research policies and evaluate studies for scientists (Porter, Cohen, Roessner, & Perreault, 2007).
Knowledge flow across various scientific disciplines normally starts from knowledge chains generated by the interactions between scientists, and becomes a knowledge flow network with the extension of the flow boundary. With the collision of knowledge and emergence of new knowledge, a knowledge field will gradually form across various scientific disciplines to accelerate innovation (van Harmelen et al., 2012). This paper investigates the evolution of the structure of an interdisciplinary knowledge flow network over different periods. Furthermore, in knowledge flow, besides the heterogeneity of knowledge distributions in different geographical locations (Gertler & Wolfe, 2006; Wu, 2012), the distributions of knowledge flowability in different scientific disciplines are also heterogeneous. The diversity and heterogeneity in science and technology is regarded as a main factor to promote innovation (Stirling, 2007). Therefore, using the dataset of scientific funding, the present paper also studies the heterogeneity of knowledge distribution in different scientific disciplines.

In previous studies, datasets used to empirically study knowledge flow across different disciplines are mainly bibliometric (Wagner et al., 2011), including citations, co-occurrences, and co-authors. In citations, cited studies have their own scientific disciplines, which can be used to calculate the disciplinary diversity by Citations Outside Category (COC) (Porter & Chubin, 1985), Brillouin’s Index (Steele, 2000), and Network Coherence (Rafols & Meyer, 2010). Interdisciplinary knowledge flow is also studied at the journal-journal citation level (Leydesdorff, Rafols, & Chen, 2013). Compared with citations, co-occurrences mainly refer to the relation among subjects or keywords. In this aspect, the basic method is co-words analysis (Gautam & Yanagiya, 2012). Using this method, some meaningful words with high frequencies are drawn out and mapped to scientific disciplines. Similar to co-occurrences, co-author relationships are generated through the collaboration among scientists. With various disciplinary background, scientists need to exchange and create knowledge together. In previous studies, researchers use literature or patents dataset to study interdisciplinary knowledge flow (Chang, Lai, & Chang, 2009; Hu, 2009).

In this paper, we intend to explore the knowledge flow in interdisciplinary research using scientific funding dataset. The relationship between scientific disciplines, especially the creation of interdisciplinary knowledge, is strongly influenced by national funding agencies (Lowe & Phillipson, 2009; Lyall et al., 2013). We propose a new measure to quantify interdisciplinary knowledge flowability. In one hand, knowledge flowability considers the cognitive distance among disciplines based on the disciplinary application codes (DAC) tree used by NSFC (Benner & Waldfogel, 2008). In the other hand, the calculation of knowledge flowability also considers the co-occurrence of scientist’s interdisciplinary applications, which represent their initial motivation of knowledge flow among their corresponding scientific disciplines. Knowledge flowability is used to weight the edges of interdisciplinary knowledge network and investigated through the analysis of the interdisciplinary knowledge network.

Furthermore, with the development of modern science and technology, Interdisciplinary research play more and more important role. Scientists themselves need interdisciplinary knowledge and also seek a variety of knowledge from other individuals through interdisciplinary collaborations (Carayol & Thi, 2005). In the application of scientific funding, they also pursue interdisciplinary applications to be successful in academic competitions. Scientific interdisciplinarity, i.e. the diversity of individual scientific disciplines, might also be
an important factor to affect individual sponsored funding. However, to the best of our knowledge, it is still unclear how scientific interdisciplinarity affects the individual sponsored funding. Therefore, in this paper, we will also try to understand the impact of scientific interdisciplinarity on the performance of scientists to be sponsored. Using the regression methods, it is found that the diversity of knowledge in terms of scientific disciplines significantly influence the number of sponsored projects and the total sponsored money at the individual scientists’ level.

**Dataset**

We collect the raw dataset from 1999 to 2013 from the ISIS system, which is an Internet-based Science Information System of NSFC (National Natural Science Foundation of China), including 193,517 records in total. Each record contains project ID, the title, project leader’s name, the amount of sponsored money, approval year, disciplinary application code, etc. Disciplinary application code, called DAC in short, indicates which discipline a project belongs to. Individuals are thus required to select the suitable DAC, which represents the specializations in their own disciplines, when they apply for scientific funding. The set of DACs has a tree structure, in which the top nodes are scientific departments, the children nodes of scientific departments are research areas, the children nodes of research area are research fields, and the children nodes of research fields are research directions. Each DAC is divided into four parts and is used to map a NSFC sponsored project into its corresponding scientific department, research area, research field and research direction, respectively. For example, the DAC of a project is A010101, which indicates that the scientific department of the project belongs to “Mathematical and Physical Science”, its scientific area is “Mathematical Theory of Numbers”, and its scientific direction is “Analytic Theory of Number”. In the NSFC, there are eight scientific departments from A to H, 86 first-level DACs, 981 second-level DACs and 1679 third-level DACs.

**Calculation of knowledge flowability**

In order to investigate knowledge flowability among scientific disciplines, the formulation of the calculation is proposed in the following.

Firstly, we calculate \( CS_{ij} \) to confirm the existence of knowledge flow between two disciplines and if it exists, to measure how frequent it is. Let \( P_i \) denote the set of scientists who have applied projects using \( DAC_i \) as the main application code. Similarly, let \( P_j \) denote the set of scientists who have applied projects using \( DAC_j \) as the main application code. \( CS_{ij} \) is the occurrence frequency of scientists between \( DAC_i \) and \( DAC_j \), which can be formulated as follows:

\[
CS_{ij} = |P_i \cap P_j|
\]

(1)

where \(|.|\) indicates the size of a set.

If \( CS_{ij} > 0 \), it means there are scientists who are overlapped in discipline \( i \) and \( j \). These scientists may have the ability to apply interdisciplinary knowledge. Scientists who do interdisciplinary research make the initial engine of knowledge integration and utilization
across different disciplines (Klein, 2008). Therefore, $CS_{ij} > 0$ somehow indicates there exists potential knowledge flow between discipline $i$ and $j$. The bigger $CS_{ij}$ is, the more frequent the potential knowledge flow is. Otherwise, if $CS_{ij} = 0$, interdisciplinary relation between $i$ and $j$ does not exist in the disciplines of science funding.

The name ambiguity have to be solved in any studies that use author name information in bibliometric analysis (Wagner, Horlings, Whetsell, Mattsson, & Nordqvist, 2015). In this paper, we did not use bibliometric data, and in our sponsored project dataset, attributes related to each sponsored scientists is his/ her name and affiliation. However, in the real world, the ambiguity of names hampers us to identify a person merely by his/ her name. A way of disambiguation is to simultaneously use the name and the affiliation or co-authorship. Therefore, we combine the names and affiliations of sponsored project leaders to uniquely identify each person in this paper (Wu & Ding, 2013).

Secondly, we calculate $W_{ij}$ that denotes the interdisciplinary score between discipline $i$ and $j$. Each DAC is divided into four parts, and is changed into a DAC set that includes four codes, i.e., the code of scientific departments, the code of research areas, the code of research fields and the code of research directions. Let $C_i$ and $C_j$ denote the DAC sets of $DAC_i$ and $DAC_j$, respectively. $W_{ij}$ is calculated as below,

$$W_{ij} = \begin{cases} 0, & C_i = C_j \\ 1, & |C_i \cap C_j| + 1, & C_i \neq C_j \end{cases}$$

(2)

where $|.|$ indicates the size of a set.

Based on this formulation, the set of values of $W_{ij}$ is $\{0, 1, 1/2, 1/3, 1/4\}$. Accordingly, interdisciplinary scores can be classified into five types: $W_{ij} = 0$ is superior-subordinate; $W_{ij} = 1$ is cross-scientific-departments; $W_{ij} = 1/2$ is cross-research-areas; $W_{ij} = 1/3$ is cross-research-fields; $W_{ij} = 1/4$ is cross-research-directions.

Thirdly, we integrate the above two calculations to calculate $FL_{ij}$. $W_{ij}$ is interdisciplinary score between two disciplines, measuring the cognitive distance based on the hierarchical structure of the disciplinary code tree. $CS_{ij}$ is the co-occurrence frequency of scientists between two disciplines, measuring the potential knowledge flow in disciplines of scientific funding. Let $FL_{ij}$ denote knowledge flowability between discipline $i$ and $j$. Therefore, we have

$$FL_{ij} = CS_{ij} \times W_{ij} = \begin{cases} 0, & C_i = C_j \\ \frac{|P_i \cap P_j|}{|C_i \cap C_j| + 1}, & C_i \neq C_j \end{cases}$$

(3)

where $|.|$ indicates the size of a set.

Based on the above calculation, we construct an interdisciplinary knowledge flow network. In the network, nodes are DACs and each node represents one discipline. If the value of knowledge flowability between two nodes is above zero, the edge between them is formed and the weight of the edge is $FL$. In order to investigate the evolution of network, we construct four sub-networks of four periods. Accordingly, the whole sample data is classified.
into four parts according to the approval years. The first is from 1999 to 2001 with a total of 199 records, the second is from 2002 to 2005 with a total of 3071 records, the third is from 2006 to 2009 with a total of 3459 records, and the forth is from 2010 to 2013 with a total of 5899 records.

Investigation of knowledge flow across disciplines
Evolution of knowledge flowability
To investigate the evolution of interdisciplinary knowledge flow network, we calculate ratios of DACs used in interdisciplinary applications for each period and the results are as follows. As shown in Figure 1, in the first period, the ratio of DACs involved in the interdisciplinary application is only 10.6 percent, while in the latest period, it has risen to 81.8 %. There are an increasing number of disciplines (nodes) involved in the interdisciplinary knowledge flow network.

Next, the changes of knowledge flowability are also investigated. In the result of Analysis of Variance (ANOV) shown in Figure 2, the value of Sig. is zero, indicating that the between-group (between different time periods) differences are significant. The changes of standard deviations and means of knowledge flowability among the first three periods are obvious, but from the third period to the last period, they are much more subtle.
In order to observe the changes of knowledge flowability in details, we analyze the types of interdisciplinary knowledge flow in each time period. There are four types of interdisciplinary knowledge flow in each time period: cross-research-departments, cross-research-areas, cross-research-fields, and cross-research-directions. We plot the ratios of all the four types in Figure 3. As the figure shows, no matter in which time period, the type of the cross-research-departments contributes the highest ratios in the whole. In addition, we can find that the ratios of the type of the cross-research-fields and the cross-research-directions are becoming bigger. It means that researches cross-research-fields and the cross-research-directions are more and more popular over years.

**Evolution of interdisciplinary knowledge flow network structure**

Structural characteristics in each time period reveal the evolution of knowledge flow network. In order to draw out these structural characteristics, some network indicators are calculated. Those indicators include the number of nodes, average degree, average weight, the density of network, average clustering coefficient and average path length, etc. These indicators are separated into three types. The first type is macro network indicators, including the number of nodes and density. The number of nodes reflects network scale. The more nodes the interdisciplinary knowledge flow network has, the wider the network boundary is. Density describes the aggregation of nodes. The network becomes more concentrated as the density grows greater. The second type of indicators shows the characteristics of one node or one edge from a micro view. They are average degree and average weight. The former one indicates the width of connection and the latter one explains the depth of connection. The last type of indicators illustrates which type the network belongs to. We can determine whether a network belongs to a random network, regular network or small world network through the comparison of average clustering coefficient and average path length. Comparatively, the average path length of random network is shortest and its average clustering coefficient is comparatively lowest, while the average path length of a regular network is longest and its average clustering coefficient is highest. The small world network has the average path length and average clustering coefficient between random network and regular network (Watts & Strogatz, 1998).
Table 1. Indicators of knowledge flow network in each time period

<table>
<thead>
<tr>
<th>Indicator types</th>
<th>Indicators</th>
<th>Time Periods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network level</td>
<td>Node</td>
<td>291</td>
</tr>
<tr>
<td></td>
<td>Density</td>
<td>0.005</td>
</tr>
<tr>
<td>Node level</td>
<td>AVG_Degree</td>
<td>1.37</td>
</tr>
<tr>
<td></td>
<td>AVG_Weight</td>
<td>0.72</td>
</tr>
<tr>
<td>Network type</td>
<td>AVG_Clustering_Coefficient(CC)</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>ACC of Random graph(CCR)</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>AVG_Path_length(PL)</td>
<td>2.14</td>
</tr>
<tr>
<td></td>
<td>APL of Random graph(PLR)</td>
<td>18.02</td>
</tr>
<tr>
<td></td>
<td>$S=(CC/CCR)/(PL/PLR)$</td>
<td>218.93</td>
</tr>
</tbody>
</table>

The structural characteristics of knowledge flow network in each time period are shown in the Table 3. With the increase of nodes, the density declines from 0.005 to 0.002. It indicates all the networks are loose. Meanwhile, with the expansion of the network size, there is no significant increase or decrease of the network cohesion indicated by the average clustering coefficient. The average degree can represent the width of connections between disciplines, and the average weight represents the depth. Overall, the width of connections between two disciplines in the network continues to increase, but the depth fluctuates around 0.7. In brief, the addition of nodes cannot deepen disciplinary connections.

At last, we analyze the type of network by comparing average clustering coefficients (CC) and average path lengths (PL) with those (CCR,PLR) of random networks generated with the same nodes(n) and the same mean links per node (k value). We use the following formula to estimate CCR and PLR: CCR~k/n, PLR~lnn/lnk (Uzzi, Amaral, & Reed-Tsochas, 2007). Typically, there are two ways to compare small-world networks to a random network. One is that PL>PLR and CC>>CCR (Uzzi et al., 2007; Watts & Strogatz, 1998). The other is a more quantitative definition of these relationships. We set $\gamma=CC/CCR$ and $\lambda=PL/PLR$. Then, we have the following equation: $S=\gamma/\lambda$. If the equation satisfies the condition: $S>>1$, the small-world exists (Davis, Yoo, & Baker, 2003; Uzzi et al., 2007). In this paper, all networks satisfy the condition of $S>>1$. Thus, we can judge that the interdisciplinary knowledge flow network is a small-world network. It means some disciplines are related to each other much more closely and the possibility of interdisciplinary collaboration and innovation due to interdisciplinary knowledge flow is considerably high.

**Heterogeneity of knowledge flowability**

Knowledge flow is heterogeneous among different disciplines. Knowledge flow among some disciplines is rather easier, because: firstly, these disciplines are somehow related; secondly, several basic disciplines whose knowledge is more inclusive and have wide influences on other disciplines; thirdly, some disciplines develop rapidly and need to absorb knowledge from others. In this part, we will explore the heterogeneity of knowledge flow across scientific departments.

In order to verify the heterogeneity of knowledge flow across scientific departments, we test the between-group differences with Analysis of Variance (ANOVA). Here, the grouping
variable is scientific departments. As shown in Figure 6, the value of Sig. is zero, which indicates that there are very highly significant differences of knowledge flowability among different scientific departments. In addition, we plot standard deviations (Std.Devs) and means of knowledge flowability in each department. Except for department D (Earth sciences), all the means are centered on 1 and all Std.Devs locate in [1, 2]. It reveals that the knowledge flowability between two disciplines in department D (Earth sciences) is more concentrated inside the scientific department than others.

![Graph showing knowledge flowability across different departments](image)

**Table 1:**

<table>
<thead>
<tr>
<th>Grouping variable</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Department</td>
<td>1165.174</td>
<td>7</td>
<td>166.453</td>
<td>70.123</td>
<td>.000</td>
</tr>
</tbody>
</table>

**Figure 4.** Comparison of knowledge flowability across scientific departments

To further compare knowledge flowability among different departments, we analyze the types of interdisciplinary knowledge flow in each department according to interdisciplinary types. There are three interdisciplinary types in one department: cross-research-areas, cross-research-fields, cross-research-directions. The ratios of all these three types in department A to H are shown in Figure 7. The type of the cross-research-areas contributes the most to the whole knowledge flow. It's worth to note that it happens in departments H, G, F, D, C, B, but not in A and E where the main interdisciplinary type is cross-research-fields rather than cross-research-areas. This means that the proportions of these three interdisciplinary types are much more balanced in scientific department A and E.

![Graph showing interdisciplinary knowledge flow types](image)

**Figure 5.** Types of interdisciplinary knowledge flow in each department
After the above comparisons, it can be confirmed that there exists the heterogeneity of knowledge flow across scientific departments. The knowledge flow differs dramatically in different scientific departments.

**Distribution of knowledge flowability**

According to the distribution of knowledge flowability, the heterogeneity of knowledge flow can be investigated directly. Therefore, we draw out the curves of knowledge flowability distributions from 1999 to 2013, as shown in Figure 8. There are four colors in the figure and one color represents one time period, respectively. The horizontal axis is the value of knowledge flowability between two disciplines, and the vertical axis is the frequency. It is found that apart from the first one whose range is only three years, the remaining three curves are very similar in shapes. The similarities appear significantly in two respects. First, there are three peaks in every period and the peaks locate in the same horizontal position. Second, besides peaks, the curves are extremely close to the horizontal axis to hold long tails (we cut off parts of the tails in the figure).

![Figure 8. Distribution of knowledge flowability](image)

Based on the above observation, we wonder if there exist some profound laws to follow or not. To further explore the potential law, some parts of knowledge flowability, whose value is 1/3, 1/2 or 1, are picked out and summed up in ratios. Knowledge flowability with these three values counts almost 80% of the total. In general, the distribution of knowledge flowability obeys the law of 80/20.

**Investigation of knowledge flow across scientists**

In order to investigate the knowledge flow across scientists in scientific funding, this section uses negative binomial regression and ordinary least squares regression to investigate the impact of individual disciplines diversity on the number of sponsored projects ($NumProj$ is...
the dependent variable in Model 1 and Model 2) and the sponsored money (SponMoney is the dependent variable in Model 3 and Model 4), respectively. In the regressions, we also control the total number of sponsored projects in the scientific department that the sponsored scientist belongs to (DeptNumProjs), the scientist’s experience that is the total years from his/her first sponsored project in NSFC (Experience), the prestige of the institution that the sponsored scientist works in (Prestige), the total number of projects that the institution of the sponsored scientist has sponsored (InsNumProjs). The main predictors are the diversity of disciplines (Diversity), which is first calculated according to Formula 2 for all the pairs of the DACs of his/her sponsored projects and then average the values of each pair of DACs.

Table 2. Diversity and innovation performance of individuals.

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>1.052***</td>
<td>1.042***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeptNumProjs</td>
<td>-0.001****</td>
<td>-0.002****</td>
<td>-0.002***</td>
<td>-0.002***</td>
</tr>
<tr>
<td>Experience</td>
<td>0.062***</td>
<td>0.053****</td>
<td>0.023***</td>
<td>0.104***</td>
</tr>
<tr>
<td>Prestige</td>
<td>0.015***</td>
<td>0.008****</td>
<td>0.017***</td>
<td>0.007***</td>
</tr>
<tr>
<td>InsNumProjs</td>
<td>0.009***</td>
<td>0.011***</td>
<td>0.026***</td>
<td>0.028***</td>
</tr>
<tr>
<td>Predictors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diversity</td>
<td></td>
<td></td>
<td>0.166***</td>
<td></td>
</tr>
<tr>
<td>Diversity squared</td>
<td></td>
<td></td>
<td>0.012***</td>
<td></td>
</tr>
<tr>
<td>Log pseudolikelihood</td>
<td>-59460.238</td>
<td>-58256.182</td>
<td>R2 0.0420</td>
<td>0.1337</td>
</tr>
<tr>
<td>Pseudo R2</td>
<td>0.0401</td>
<td>0.0596</td>
<td>Adjusted R2 0.0419</td>
<td>0.1335</td>
</tr>
</tbody>
</table>

* p< 0.05, ** p< 0.01, ***p< 0.001

Through the regressions, it is found that Diversity and Diversity squared in the knowledge flow across scientists both positively affect the scientists’ innovation performance in term of NumProj and SponMoney, respectively. This main finding is interesting to indicate that a good level of interdisciplinary knowledge flow is very important for a scientist to be successfully sponsored by scientific funding.

Conclusions

One of the most important characteristics of modern science is interdisciplinary (Porter & Rafols, 2009). About 46.8% of the 5500 existing disciplines are interdisciplinary and new inter-disciplines are emerging as well (Research, 2005). Under this situation, the NSFC encourage interdisciplinary applications through a series of policies such as disciplines union funding support. By the end of 2013, more than 59% applicants have ever changed disciplinary application code during application (Wu et al., 2015). The disciplinary application code plays a key role in the selection of review experts during the peer review. It is necessary to understand the relationships among disciplines to to well guide the funding applications.

This paper investigates the interdisciplinary knowledge flow across disciplines and scientists based on disciplinary application codes of scientific funding. The interdisciplinary
knowledge flow network is made up of disciplines as nodes and their relationships as links. The nodes of disciplines correspond to disciplinary application codes. The links among disciplines represent the co-occurrences of disciplinary application codes (DACs). Through analyzing knowledge flow networks among scientific disciplines and investigating the impact of scientific interdisciplinarity on the performance of scientists, the findings have some implications from a scientific policy perspective, and can also give some suggestions to the policy interventions in research practices.

Therefore, this paper studies the interdisciplinary knowledge flow across disciplines and scientists, and gives some suggestions to policy makers of scientific research. Nevertheless, our research is not without limitations. Although we have solved the ambiguity of applicant names, the ambiguity also exists among DACs in the dataset because it cannot be completely solved and the discipline classification scheme has not been perfected with the development of interdisciplinary trend in science and technology (Huutoniemi, 2012). In the future work, the impact of network effects on the total sponsored money and the number of sponsored projects in a certain scientific discipline can be studied as well.

Acknowledgements
This work was supported by grants from National Natural Science Foundation of China (grant no.: 71373194) and the China Postdoctoral Science Foundation (grant no.: 2014T70143).

References
Leydesdorff, L., Rafols, I., & Chen, C. 2013. Interactive overlays of journals and the measurement of


The association between university research indicators and success rates in the European Framework Programmes

Fredrik Niclas Piro· Lisa Scordato· Dag W. Aksnes
fredrik.piro@nifu.no lisa.scordato@nifu.no dag.w.aksnes@nifu.no
Nordic Institute for Studies on Innovation, Education and Research (Norway)

Abstract

In this study, we investigate which university research indicators (size, citation impact, industry collaboration and reputation) are associated with success in the European Framework Programmes for Research and Innovation (the Seventh Framework Programme and Horizon 2020). Based on a unique dataset involving all applicants to and project participants in EU projects, we have calculated success rates for all universities in the world that have applied for EU funding in the period 2007-2015. Linked to results from the Shanghai Ranking and to publication indicators from the Leiden Ranking, we demonstrate how such university research indicators are correlated with success rates in various EU programmes. Our analysis shows that highly cited papers have the highest correlation with success rates.

Conference Topic
Indicators; Science policy and research assessment; University policy and institutional rankings

Introduction

There are large differences in how institutions perform in university rankings. The most influential rankings allegedly reflect various aspects of university performance, such as the reputation or research quality of an institution; to a lesser degree the teaching quality. Similarly, there are large differences across universities in their success when applying for external funding. Although there are several limitations attached to university rankings, one would expect that there is a certain correlation between the two measures: The higher ranked institutions (by high quality research or high international reputation) would have higher success rates in terms of obtaining research funding. In this study, we attempt to investigate to what extent the institutions’ reputation and publication and citation patterns as reflected in university rankings are associated with their success rates in the European Union’s Framework Programmes for Research and Innovation (EU FPs). There is little evidence about this correlation, because the data that is available – and the studies using these data – have relied upon participation data in FPs, and not on application data, thereby not taking the ratio between number of submitted proposals and number of granted projects (i.e. success rates) into account.

Some scholars have studied patterns and determinants of R&D collaborations using datasets on participation in EU FPs (e.g. Roediger-Schluga & Barber 2008; Paier & Scherngell 2011; Enger & Castellaci 2016). In a recent study by Lepori et al. (2015), it was demonstrated that in a sample of 2,235 European higher education institutions, the level of participation in EU FP funded projects was strongly associated with organizational characteristics, particularly with size and reputation. The finding is in accordance with studies that have shown that acquisition of research funds also at the individual level is more strongly correlated with the reputation of the research applicant than with the judgement of the proposal quality in the selection process (cf. Hamann, 2016). With respect to R&D collaboration, Lepori et al. (2015, p. 2153) argue that: “higher-reputed researchers and organizations will be sought to a greater extent as research
partners and, therefore, move to the center of the network”, but such an assumption – at the university level – does not take into account that the highly reputed institutions may be at the centre, i.e. participating in many collaborative projects, also because they simply submit many proposals, reflecting their large institutional size. Piro, Scordato & Aksnes (2016) investigated all EU FP consortia in EUs Seventh Framework Programme and Horizon 2020, involving participation from six European countries and concluded that a strong determinant for success in FPs was previous participation in FP projects and a large network of European partners from past and ongoing projects, indicating that institutions that are already central in the European research network will become even more central as they are attractive partners to engage with for other institutions, thus strengthening the ‘Matthew effect’ of FP participation.

Research questions
The aim of this study is to examine the correlation between universities’ success rates in EU’s Seventh Framework Programme (FP7) and Horizon 2020 (H2020) and their size (measured as publication volume), research impact (measured as citation impact), industry orientation (measured by co-authored publications with industrial partners) or with reputation (measured as position in global university rankings). The research questions of our paper have the universities’ success rate in EU FPs at their centre, i.e. the ratio between an institution’s number of submitted proposals and its number of funded projects. We examine whether this measure is differently associated with the four university research indicators described above. The study is carried out in two steps. First, we make a correlation analysis of the relationship between success rates and university indicators. Second, we carry out fine-tuned analyses, where the following four correlations are investigated:

Across thematic areas in the FPs (1). The structure of FP7 and H2020 is different, with different organizing of thematic areas/pillars, emphasis of action instruments, demands for technology readiness levels (TRLs), demands for inclusion of industry and end-users in consortia, varying degrees of basic and applied research, etc. Altogether, considering a university’s total success rate for an entire FP may not be very useful as a university may be highly successful within the thematic area closest to the main research areas of the university, while may still be participating in proposals to a range of more distant topics that in sum may exceed the volume of the former. In some thematic areas of the FPs, academic qualifications and merits may play a larger role in the review process, compared to other areas where the actions are more aimed at private for profit companies, thus making the universities’ role less crucial.

University as coordinator vs. ordinary partner (2). The coordinator is the driver of the project, and in a consortium where the number of participants may vary from 2-3 to more than a hundred, it seems plausible that the reputation or impact of the coordinating university will be more important than those of the other universities involved in the consortium. In other consortia, which may be coordinated by e.g. private companies, it is an open question whether the inclusion of a highly reputed university among the partners will strengthen the proposal’s prospects for funding. Alternatively, the review process may be so holistic in its consortia assessments that no special attention is given to the coordinating institution - rather the consortia is being judged based on the totality of all involved partners.

The European Research Council (ERC) vs. collaborative programmes (3). One distinction that can be made between different elements of the FPs is that the ERC in contrast to other instruments is less focused on research collaboration. Decisions on ERC funding are made based on criteria related to scientific excellence only. This criterion is applied for assessing both the proposed research and the Principal Investigator. The main purpose of ERC is to support
investigator driven, frontier research, across all fields. ERC belongs (now) formally to the first pillar of the H2020 programme promoting ‘Scientific Excellence’ alongside e.g. EU’s mobility programme Marie Skłodowska-Curie actions (MSCA) and the programme Future and Emerging Technologies (FET), aimed at initiating radical new technology through unexplored collaborations between advanced multidisciplinary science and cutting-edge engineering. In the review of the applications to the ERC, the panels have used three criterions: 1) scientific quality of the applicant, 2) quality of the research project proposed by the applicant, 3) assessment of the scientific environment at the institution where the applicant’s research would be conducted. In later calls, only the first two criterions have been used. Such criterions are different from other programmes, and may therefore lead to different importance of universities’ citation impact and reputation in the review process. For the other programmes under H2020, the general rule is that experts are to assess proposals on the basis of three criteria: excellence, impact, and quality and efficiency of the implementation (Langfeldt & Scordato, 2016).

Geographical origin: European vs. non-European countries (4). It is fully possible that e.g. citation indicators are more associated with overall success rates compared to results from university rankings in some programmes, while the opposite may be the case in some thematic areas or for universities from some specific geographical areas. For example: the citation impact may be important in explaining differences among European universities, while rank position in a global university ranking may be more associated with differences in success rates between universities from Asia and South-America. The geographic dimension is important because many of the most successful universities in the world, according to our selected indicators, are from the United States (in particular), Canada and Asian countries. These countries seldom contribute as coordinators and their total participation in the FPs is limited compared to European institutions, thus making their centrality in the FP networks less pronounced.

Methods and data sources
This study draws upon three main data sources: 1) data about EU FP proposals and applications from the European Commission’s data warehouse ECORDA, 2) data from the Leiden Ranking produced by Centre for Science and Technology Studies (CWTS) at Leiden University, and 3) data from the Academic Ranking of World Universities (ARWU) from the Shanghai Jiao Tong University, commonly referred to as the Shanghai Ranking. Three of our university indicators are taken from the Leiden Ranking: Number of fractionalized publications (PUBS), share of publications co-authored with industry (INDUSTRY), and percentage of the world’s ten per cent most cited publications (TOP10%) (see Waltman et al. 2012). The fourth indicator is the universities’ mean rank position in the Shanghai Ranking (ARWU).

ECORDA data
The main data source in this study is ECORDA data, covering FP7 and the early phase of H2020. We have used the November 2015 edition, which means that our FP7 data are complete, whereas the analysis of H2020 is restricted to the early results of that framework programme. In ECORDA, the institutional affiliations of applicants are not standardized. It is therefore not possible to use the data for calculating success rates without going through the process of standardizing the institutional names. The novelty of our study is the build-up of a completely new database for participation in both EU proposals and projects. At Nordic Institute for Studies in Innovation, Research and Education (NIFU), all institutional addresses in EU’s application and project databases of FP7 and H2020 have been standardized (approximately 1.1 million institution names). To the best of our knowledge, such a standardized data file including both applicants and grant recipients does not exist anywhere else. This has enabled us to extract data
about all institutions’ total volume of applications and projects, thereby making it possible to

In targeting universities for the present study, we excluded all university hospitals, and left out
continuing education, training centres, etc., thus focusing the analysis on ‘pure’ universities and
university colleges. All institutions in ECORDA with a research and/or teaching activity has
been investigated in detail to ensure that only universities are brought into the analyses. Based
on the number of applications and project participations, we calculate each university’s success
rate in each FP programme separately, and overall for FP7 and H2020 combined.

Number of publications from the Leiden Ranking (PUBS)
The Leiden Ranking includes data on the publication volumes of the world’s largest
universities. Lepori et al. (2015) investigated whether the number of participations in FPs is
expected to increase with the research capacity of the organization. They postulated that
organizations with more research capacity have more research teams (possibly also covering
more research topics), and therefore engage in more collaborations. Size of an institution may
also determine whether or not the institution has developed an administrative capacity that may
facilitate participation in FP applications and projects, which is often seen as crucial for
researchers (Piro, Scordato & Aksnes, 2016). Since we do not have institutional data for the
universities – we use the number of scientific publications (PUBS) as a size indicator. All
universities in the Leiden Ranking have been given a value representing their mean number of
fractionalized scientific publications during the years 2011-2014 (the Leiden Ranking does not
provide numbers further back in time).

Top 10% publications from the Leiden Ranking (TOP10%)
In the Leiden Ranking, there are two main citation indicators: mean normalized citation score
(MNCS) and percentage of highly cited papers, i.e. top 10 per cent publications (TOP10%).
The latter indicator measures the proportion of a university’s publications that, compared with
other publications in the same field and from the same year, belong to the ten per cent most
frequently cited in the whole world. Since MNCS and TOP10% are strongly correlated we
chose TOP10% as this reflects the proportion of a university’s publications with very high
impact, compared to MNCS which provides an average of all research at a university, from
uncited works to extremely cited works. We refer to Waltman et al. (2012) for a more detailed
discussion of the MNCS and TOP10% indicator.

Share of co-authored publications with industry from the Leiden Ranking (INDUSTRY)
The Leiden Ranking provides the proportion of a university’s publications that have been co-
authored with one or more industrial partners (INDUSTRY). We use this as a proxy for how
business/industry oriented a university is.

University ranking data from the Shanghai Ranking (ARWU)
The relevance and quality of university rankings are highly disputable (Piro & Sivertsen, 2016),
but they are well-known and may be used to investigate the ‘quality’ of the institutions the
applications are affiliated with. Despite the various limitations of the university rankings’
indicators and methods underlying them (Docampo and Cram 2015; Safón 2013), few would
disagree that the universities ranked at the top in these rankings are prestigious and leading
institutions. We may claim that the rankings are able to identify the best universities of the
world, but fail in discriminating between ‘regular’ universities. Here, we compare the
universities based on their ranking in the Shanghai Ranking (ARWU), providing the
universities with a mean value of their rank position in the years 2003-2014. A large portion of
the universities’ ARWU scores are derived from their publication output and a citation indicator (cf. PUBS and TOP10%). Nevertheless, it is important to note that the methodology is very different. ARWU measures whole counts of publications, while we measure fractionalized publications; ARWU measures highly cited researchers (based on individuals featuring on Thomson Reuters’ list of highly cited researchers), while we measure highly cited publications.

Programmes studied in EU FPs

Although some of the FP7 programmes are possible to compare with H2020 programmes, others are not. Therefore, the only consistent way of comparing success rates is to do so for all programmes separately. We have restricted our study to seven main programmes of the FPs: (1) The European Research Council (ERC) and (2) Marie Curie Actions (MCA in FP7)/Marie Skłodowska-Curie Actions (MSCA in H2020) are comparable across FP7 and H2020 and therefore features as four distinct programmes. (3) Cooperation (FP7): Energy; Environment (incl. Climate Change); Food, Agriculture & Biotechnology; Health; Information and Communication Technologies; Joint Technology Initiatives; Nanosciences, Nanotechnologies; Materials and new Production Technologies; Security; Socio-economic Sciences & Humanities; Space; Transport (incl. Aeronautics). (4) Societal Challenges (H2020): Climate action, environment, resource efficiency and raw materials; Europe in a changing world – inclusive, innovative and reflective Societies; Food security, sustainable agriculture and forestry, marine and maritime and inland water research; Secure societies – Protecting freedom and security of Europe and its citizens; Secure, clean and efficient energy; Smart, green and integrated transport; Health, demographic change and wellbeing. (5) Industrial Leadership (H2020): Access to risk finance; Advanced manufacturing processing; Advanced materials, Biotechnology; Cross-theme; Nanotechnologies, advanced materials and production; Space; Information and communication technologies. (6) Science with and for Society (H2020): Develop the governance for the advancement of responsible research and innovation; Integrate society in science and innovation; Make scientific and technological careers attractive for young people; Promote gender equality in research and innovation. (7) Future and Emerging Technologies (H2020), which is one of four programmes under the Excellence pillar of H2020 (alongside ERC, MSCA and Research infrastructures – the latter not part of our analysis).

Study Sample

In ECORDA we have identified 4,957 higher education institutions from a total of 174 countries. Of these 697 are included in the Leiden Ranking and 482 in ARWU. Combined there are 718 universities from 49 countries that are included in one, or both, of these rankings. With more than 200 universities in total, the United States and China are highly dominating the list of universities. 270 universities are from EU member states, 32 from other European countries, 191 from North America, 20 from Latin- and South America, 168 from Asia, 28 from Australia/New Zealand, and 9 from Africa. The universities are analysed separately and in groups. For each university research indicator, we have grouped the universities in deciles by their position on the respective indicator, so that each university (in most cases) will be assigned to different categories for different indicators. In ARWU the universities are grouped in ten groups of 41-52 universities, whereas the universities in the Leiden-based indicators are grouped in nine groups of 69-76 universities. In analyses where the results are split by geographical regions, we distinguish between EU28 and associated countries to H2020 (as of January 1st 2017 includes universities from Israel, Norway, Serbia, Switzerland, Tunisia and Turkey) and the rest of the world.
Results

Table 1 shows the key values for the (deciles of) university groups by four university indicators: Mean position in the Shanghai Ranking (ARWU), mean number of fractionalized publications in the Leiden Ranking (PUBS), mean shares of publications co-authored with industry (INDUSTRY) and mean share of top 10 per cent publications in the Leiden Ranking (TOP10%). These numbers are based on overall performance in both FP7 and H2020 and do not distinguish between programmes, thematic areas and so forth. A few inconspicuous findings emerge. First, in all university groups we find universities with a success rate of zero, i.e. they have not taken part in any proposals that have been granted funding. Most of these ‘unexpected cases’ originates from the US university system, but it is important to note that their contributions to proposals is not low – in some cases US universities have participated in more than 200 proposals, resulting in no funding. Second, due to the large US contribution to the first university categories (1), it appears that the top-tiers for all three indicators are found in university groups 2 and 3, having a larger European contribution of universities. Third, the activity measures (applications and projects) of EU FP involvement, is clearly diminishing from university groups 2 and 3 and onwards, which is hardly surprising as the universities are getting smaller the closer we are approaching university group 10. Fourth, mean success rates of universities by groups do not follow a similarly easily detectible pattern. Based on ARWU, the lowest success rates are indeed found among the highest ranked universities, while the mid-ranked universities (groups 5 and 6) have the highest success rates (15 and 16 per cent respectively).

Table 1. Success rates and number of applications/projects for university groups in FP7 and H2020, across ten university groups.

<table>
<thead>
<tr>
<th>ARWU</th>
<th>N</th>
<th>Number of applications</th>
<th>Number of projects</th>
<th>Success rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>1</td>
<td>48</td>
<td>524</td>
<td>7</td>
<td>3567</td>
</tr>
<tr>
<td>2</td>
<td>48</td>
<td>674</td>
<td>13</td>
<td>2238</td>
</tr>
<tr>
<td>3</td>
<td>49</td>
<td>733</td>
<td>7</td>
<td>3534</td>
</tr>
<tr>
<td>4</td>
<td>49</td>
<td>380</td>
<td>1</td>
<td>2223</td>
</tr>
<tr>
<td>5</td>
<td>52</td>
<td>427</td>
<td>1</td>
<td>2458</td>
</tr>
<tr>
<td>6</td>
<td>48</td>
<td>469</td>
<td>1</td>
<td>2057</td>
</tr>
<tr>
<td>7</td>
<td>50</td>
<td>275</td>
<td>1</td>
<td>1378</td>
</tr>
<tr>
<td>8</td>
<td>48</td>
<td>352</td>
<td>1</td>
<td>1598</td>
</tr>
<tr>
<td>9</td>
<td>49</td>
<td>221</td>
<td>1</td>
<td>1549</td>
</tr>
<tr>
<td>10</td>
<td>41</td>
<td>291</td>
<td>1</td>
<td>1456</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PUBS</th>
<th>N</th>
<th>Number of applications</th>
<th>Number of projects</th>
<th>Success rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>1</td>
<td>69</td>
<td>465</td>
<td>3</td>
<td>3567</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>686</td>
<td>1</td>
<td>3534</td>
</tr>
<tr>
<td>3</td>
<td>69</td>
<td>580</td>
<td>3</td>
<td>2481</td>
</tr>
<tr>
<td>4</td>
<td>69</td>
<td>352</td>
<td>1</td>
<td>2062</td>
</tr>
<tr>
<td>5</td>
<td>69</td>
<td>463</td>
<td>1</td>
<td>2057</td>
</tr>
<tr>
<td>6</td>
<td>69</td>
<td>368</td>
<td>1</td>
<td>1726</td>
</tr>
<tr>
<td>7</td>
<td>69</td>
<td>230</td>
<td>1</td>
<td>1282</td>
</tr>
<tr>
<td>8</td>
<td>69</td>
<td>151</td>
<td>1</td>
<td>778</td>
</tr>
<tr>
<td>9</td>
<td>69</td>
<td>212</td>
<td>1</td>
<td>1073</td>
</tr>
<tr>
<td>10</td>
<td>76</td>
<td>125</td>
<td>1</td>
<td>782</td>
</tr>
</tbody>
</table>
In the indicator PUBS, success rate is lowest in the expected university group (10), and highest in the third and sixth university group, while for a university stratification based on TOP10% the highest success rate is in the second highest ranked group of universities.

**Correlations**

Again, without distinguishing between different types of programmes, thematic areas and so on, a correlation analysis (Pearson’s r) found no significant correlations between the university indicators and success rates. The only exception was a weak correlation between INDUSTRY and success rates.

<table>
<thead>
<tr>
<th>EU 28 and associated countries</th>
<th>PUBLS</th>
<th>INDUSTRY</th>
<th>TOP10%</th>
<th>ARWU</th>
<th>Success rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUBLS</td>
<td>1</td>
<td>.214**</td>
<td>.494**</td>
<td>.731**</td>
<td>.405**</td>
</tr>
<tr>
<td>INDUSTRY</td>
<td>1</td>
<td>.437**</td>
<td>.151*</td>
<td>.438**</td>
<td></td>
</tr>
<tr>
<td>TOP10%</td>
<td>1</td>
<td>.588**</td>
<td>.654**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARWU</td>
<td>1</td>
<td>.424**</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Non-European countries</th>
<th>PUBLS</th>
<th>INDUSTRY</th>
<th>TOP10%</th>
<th>ARWU</th>
<th>Success rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUBLS</td>
<td>1</td>
<td>.307**</td>
<td>.530**</td>
<td>.673**</td>
<td>.044</td>
</tr>
<tr>
<td>INDUSTRY</td>
<td>1</td>
<td>.435**</td>
<td>.395*</td>
<td>.076</td>
<td></td>
</tr>
<tr>
<td>TOP10%</td>
<td>1</td>
<td>.688**</td>
<td>-.137**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARWU</td>
<td>1</td>
<td>.100**</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Correlations is significant at the 0.01 level (2-tailed). * Correlation is significant at the 0.05 level (2-tailed).**
However, when splitting the results by European vs. non-European countries, the results change markedly (Table 2). Outside Europe, the correlations are weak and appear random: high shares of TOP10% are even negatively associated with success rates, while for high ARWU rank it is the opposite. We will therefore only pursue results for European countries in upcoming tables and figures. The correlations are split in Table 3 by a sample of nine main thematic areas/sub-programmes, and shown separately for coordinators and partners. In the Cooperation programme of FP7, all four indicators are significantly correlated to coordinators’ success rates, most notably the TOP10% indicator. 308 universities in the Leiden Ranking had coordinated an application submitted to the Cooperation programme. This programme does not have any direct equivalent in H2020, rather, the sub-programmes of the Cooperation programmes are now found in the pillars Industrial Leadership (e.g. ICT) and Societal Challenges (e.g. Health).

### Table 3. Correlations between key indicators and success rates across main FP7 and H2020 programmes (only significant numbers are shown) for EU28 and associated countries.

<table>
<thead>
<tr>
<th>Program</th>
<th>University role</th>
<th>PUBS</th>
<th>TOP 10%</th>
<th>INDUST.</th>
<th>ARWU</th>
<th>N Leiden</th>
<th>N ARWU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP7 Cooperation</td>
<td>Coordinators</td>
<td>.208**</td>
<td>.355**(3)</td>
<td>.223**</td>
<td>.152*</td>
<td>299</td>
<td>203</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H2020 Industrial Leadership</td>
<td>Coordinators</td>
<td>.204**(1)</td>
<td>.132*</td>
<td></td>
<td></td>
<td>249</td>
<td>177</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td>.207**</td>
<td>.143**</td>
<td></td>
<td></td>
<td>334</td>
<td>228</td>
</tr>
<tr>
<td>H2020 Societal Challenges</td>
<td>Coordinators</td>
<td>.173**</td>
<td>.224**(1)</td>
<td>.178**</td>
<td>.150*</td>
<td>281</td>
<td>195</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td>.122*</td>
<td>.272**(3)</td>
<td></td>
<td></td>
<td>298</td>
<td>202</td>
</tr>
<tr>
<td>H2020 Science with and for Society</td>
<td>Coordinators</td>
<td></td>
<td>-.303*(2)</td>
<td></td>
<td></td>
<td>60</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>201</td>
<td>146</td>
</tr>
<tr>
<td>FP7 Marie Curie</td>
<td>Coordinators</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H2020 Marie Curie</td>
<td>Coordinators</td>
<td>.131*(1)</td>
<td>.115*(2)</td>
<td>.181*</td>
<td></td>
<td>294</td>
<td>201</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>300</td>
<td>203</td>
</tr>
<tr>
<td>H2020 FET</td>
<td>Coordinators</td>
<td>.158*</td>
<td>.145*</td>
<td></td>
<td></td>
<td>216</td>
<td>157</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>286</td>
<td>199</td>
</tr>
<tr>
<td>FP7 ERC</td>
<td>Coordinators</td>
<td>.352**</td>
<td>.485**(3)</td>
<td>.287**</td>
<td>.499**(3)</td>
<td>299</td>
<td>204</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td>.115*</td>
<td>.237**(3)</td>
<td></td>
<td></td>
<td>300</td>
<td>204</td>
</tr>
<tr>
<td>H2020 ERC</td>
<td>Coordinators</td>
<td>.168**</td>
<td>.310**(3)</td>
<td>.166**</td>
<td>.238**</td>
<td>287</td>
<td>203</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td>.193**</td>
<td>.174**</td>
<td>.254**</td>
<td></td>
<td>289</td>
<td>203</td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level (2-tailed). * Correlation is significant at the 0.05 level (2-tailed). (1) Regression estimate significant at the 0.05 level (2-tailed), (2) Regression estimate significant at the 0.01 level (2-tailed) and (3) Regression estimate significant at the 0.005 level (2-tailed).

In H2020’s Industrial Leadership only INDUSTRY is correlated with success rates, while it is foremost TOP10% and PUBS that are correlated in the Societal Challenges program. In the Marie Curie mobility programmes, there is very little evidence of university indicators being correlated with success rates, while the only significant correlation found in the program Science with and for Society, is a negative correlation between ARWU and success rate, clearly indicating that this programme is not targeting the most prestigious universities.

Due to the correlations between our university research indicators, we also performed an OLS regression analysis to determine whether their significant correlations with success rates still
held. In Table 3, the significance levels of the regression estimates are indicated by (1) significant at the 0.05 level, (2) significant at the 0.01 level and (3) significant at the 0.005 level. As expected, the most consistent set of associations between university research indicators and success rates is found for ERC. It is also evident that TOP10% and ARWU are more strongly correlated to success rates than PUBS and INDUSTRY. The programme Future and Emerging Technologies (FET) under the Excellence pillar of H2020 is a special case, as it has become the programme with the lowest success rate in H2020 due to the large volume of submitted proposals. FET consortia are typically much smaller than consortia in Industrial Leadership and Societal Challenges, and it is only coordinators’ TOP10% and INDUSTRY values that are correlated to success rates. The main tendency in Table 3 is that the university indicators for coordinating universities are more related to success rates than for universities in a partner role, and that TOP10% is clearly more associated with success rates across the programmes than the other indicators.

**Prediction analysis**

The results so far have been based on correlation analyses where each university is treated as a separate unit. We now look at the ten university groups that were constructed for each university indicator. Based on mean success rates within each group we use a simple prediction test of how many correctly predicted rank values each university group has for the indicators. In a selection of ten groups, if a university group with e.g. the third highest ARWU score also has the third highest success rate, then the ARWU score is 100 per cent correct predicted for this group.

**Table 4. Percentage of correctly estimated university group values for EU28 and associated countries.**

<table>
<thead>
<tr>
<th>Program</th>
<th>University role</th>
<th>PUBS</th>
<th>TOP10%</th>
<th>INDUSTRY</th>
<th>ARWU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP7 Cooperation</td>
<td>Coordinators</td>
<td>72 %</td>
<td>92 %</td>
<td>76 %</td>
<td>60 %</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td>76 %</td>
<td>84 %</td>
<td>84 %</td>
<td>76 %</td>
</tr>
<tr>
<td>H2020 Industrial Leadership</td>
<td>Coordinators</td>
<td>44 %</td>
<td>60 %</td>
<td>64 %</td>
<td>36 %</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td>56 %</td>
<td>36 %</td>
<td>68 %</td>
<td>44 %</td>
</tr>
<tr>
<td>H2020 Societal Challenges</td>
<td>Coordinators</td>
<td>72 %</td>
<td>76 %</td>
<td>64 %</td>
<td>64 %</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td>64 %</td>
<td>68 %</td>
<td>72 %</td>
<td>52 %</td>
</tr>
<tr>
<td>H2020 Science with and for Society</td>
<td>Coordinators</td>
<td>32 %</td>
<td>24 %</td>
<td>52 %</td>
<td>12 %</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td>40 %</td>
<td>52 %</td>
<td>56 %</td>
<td>36 %</td>
</tr>
<tr>
<td>FP7 Marie Curie</td>
<td>Coordinators</td>
<td>72 %</td>
<td>60 %</td>
<td>52 %</td>
<td>72 %</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td>32 %</td>
<td>32 %</td>
<td>40 %</td>
<td>48 %</td>
</tr>
<tr>
<td>H2020 Marie Curie</td>
<td>Coordinators</td>
<td>76 %</td>
<td>72 %</td>
<td>56 %</td>
<td>80 %</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td>32 %</td>
<td>32 %</td>
<td>64 %</td>
<td>20 %</td>
</tr>
<tr>
<td>H2020 FET</td>
<td>Coordinators</td>
<td>60 %</td>
<td>92 %</td>
<td>56 %</td>
<td>60 %</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td>40 %</td>
<td>24 %</td>
<td>28 %</td>
<td>48 %</td>
</tr>
<tr>
<td>FP7 ERC</td>
<td>Coordinators</td>
<td>96 %</td>
<td>100 %</td>
<td>88 %</td>
<td>96 %</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td>68 %</td>
<td>80 %</td>
<td>68 %</td>
<td>56 %</td>
</tr>
<tr>
<td>H2020 ERC</td>
<td>Coordinators</td>
<td>60 %</td>
<td>88 %</td>
<td>72 %</td>
<td>76 %</td>
</tr>
<tr>
<td></td>
<td>Partners</td>
<td>56 %</td>
<td>64 %</td>
<td>44 %</td>
<td>68 %</td>
</tr>
</tbody>
</table>

If the university group with the lowest ARWU scores (that is, group 10) has the highest success rate, then the ARWU rank of this group completely fails to predict the success rate. The
prediction value of each indicator is then the sum of all university groups’ deviations from a correctly predicted success rate rank. The first observation, cf. correlations in Table 4, is the very low predictive power of university groups for partners’ success rates. It is practically non-existing in many programmes. The main result is that when universities are coordinating applications, the university group values correspond with success rates. ERC, FP7 Cooperation and H2020 Societal Challenges may all be characterized as programmes with good overall association between university research indicators and success rates. In the FET program it is striking how the TOP10% indicator is correlated to success rates. In other programmes such as Marie Curie and Science with and for Society there is no evidence at all for such associations. TOP10% values of coordinators generally stand out as most linked to success rates.

We illustrate these findings by analysing the coordinating universities in FP7 Cooperation (Figure 1). We find that for all four indicators, except INDUSTRY, the success rates are highest in university groups 1 (i.e. the groups consisting of universities with the highest indicator values). For the indicator PUBS, there is a clear tendency towards declining success rates as we move from group 1 to group 10, but from the third group and onwards, it levels out. For ARWU values, there are no systematic patterns at all. However, when we study TOP10%, there is a systematic decline from university group 1 to university group 10. The success rate of the university group with the lowest TOP10% values is less than 20 per cent of what the universities with the highest shares of highly cited papers have.

**Discussion and conclusions**

The findings of our study implicate that analysing publication, citation and reputation indicators in relation to success rates in the European Framework Programmes (EU FPs) do make sense, i.e. the indicators that we have used are associated with success rates in plausible manners, indicating that these associations are not random or due to statistical noise. Our analysis
demonstrates that universities’ success rates in EU FPs are less size- and reputation dependent than what might be expected, and more related to citation impact.

Previous studies have shown that the large, prestigious universities dominate EU FP participation. This study provides new insight into characterizing European research as ‘elite-driven’. The basis for the analysis is the yet unexplored possibility that large and prestigious universities are more involved in EU FPs simply because they are either highly research intensive and/or has a size allowing them to engage in more proposals. Our standardization of ECORDA has, for the first time, enabled comparative analysis of success rates for all universities in the world. The study thus adds knowledge about these universities’ relative strong contribution to EU FP participations. The analysis point at the following results for the four dimensions studied in this paper:

**University as coordinator vs. partner (1):** The success rates for universities differ and depend on whether they have coordinated or contributed as partner in a proposal. The university research indicators related to the coordinating university’s size, industrial collaboration, citation impact and global reputation appear to be most important for success rates. An explanation may be that the coordinator characteristics are given more weight in the review process of proposals than partners’ characteristics in some programmes. In other programmes, it is the consortia itself that is reviewed based on the partners’ complementarity in fulfilling the aims of the project. Here, we may postulate that ‘strong’ universities attract stronger partners to their consortia, thus making the characteristics of the coordinating university associated with the partners’ characteristics.

**Across thematic programmes (2):** The correlations of university research indicators and success rates differ between the EU FP programmes, e.g. the TOP10% indicator stands out as the ‘ERC indicator’, the INDUSTRY indicator as the ‘Industrial Leadership indicator’, ARWU’s negative association in the Science with and for society programme illustrates how this program is lesser oriented towards research excellence than other EU programmes. These findings are not surprising as the indicators may have varying relevance across programmes and instruments. For example, the Marie Sklodowska Curie actions aims at targeting young researchers, stimulating mobility and cooperation, and build effective cooperation between science and society; aspects that are unlikely to be reflected through the selected indicators. The university research indicator that is most strongly associated with success rates in EU FPs is the share of highly cited papers, that is, papers that belong to the ten per cent most frequently cited publications. This indicator is strongly correlated to the mean citation rate of publications, but not so with a university’s total number of publications, its share of co-authored publications with industry or with its position in ARWU. It is not foremost large or highly ranked universities that obtain high success rates in the FPs, rather it is the highly cited universities.

**European vs. non-European countries (3):** One possible mechanism behind the often very low success rates of the large non-European universities may be that they are peripheral actors in the EU FP network. Piro, Scordato & Aksnes (2016) defined ‘centrality’ as the sum of unique partners an institution had been working with in funded EU-projects, and found a strong relationship between centrality and success rates (also between the number of project participations and success rates, clearly providing support to the ‘Matthew effect’ in EU FPs). Universities that are on the inside of the EU FP networks benefit from their institutional reputation, their know-how on how to write proposals and their networks that they can draw upon. The importance of strengthening one’s degree of centrality in the European research network, has even been highlighted by the European Commission (2015a, p.112) as an
important driver towards success, recommending that institutions should make efforts in increasing their networks. However, the landscape of FP partners seems fixed. The growth in FP funding over time has not been followed by a corresponding increase in the number of units that receive funding (European Commission, 2015b). Our results certainly do not support any claims that bringing e.g. highly prestigious US universities into a consortium is beneficial for the prospects of funding. In fact, bringing in a prestigious university into a consortium only seems a proper strategy when the university is already a key player in FP research. Although it is possible for partners from outside of the EU to coordinate projects, such coordination is almost non-existing, and our results have clearly demonstrated that the university research indicators are almost exclusively associated with success rates for universities in a coordinating role, hence the lack of significant correlations for universities from outside of Europe.

**ERC vs. collaborative programmes (4):** There are large differences in various EU programmes in how university research indicators are related to success rates. The most distinct finding was between ERC other thematic areas. As expected, university indicators measuring scientific excellence was more accentuated in ERC, but a startling discovery was how the split-up of FP7’s Cooperation program to new programmes in H2020 led to much smaller associations between these indicators and success rates. Furthermore, industry collaboration became more evident as a factor in understanding success rates. We believe this must be seen in light of the reorientation in H2020 towards commercialization and societal impact.

**References**


From academia to citizenry. Study of the flow of scientific information from R&D projects to scientific journals and social networks

Daniela De Filippo¹,²,³ and Antonio Eleazar Serrano-López¹,²,³

dfilippo@bib.uc3m.es; aeserran@bib.uc3m.es

¹Research Institute for Higher Education and Science (INAECU) (UC3M-UAM) (Spain)
²Laboratory for Metric Information Studies (LEMI), Department of Library and Information Science, Carlos III University of Madrid (Spain)
³Associated unit Institute of Philosophy (CSIC), LEMI (UC3M), (Spain)

Abstract
This article analyses the flow of scientific information across several domains. The point of departure was the collection of data on the characteristics of European projects on energy savings funded under the Seventh Framework Programme. After bibliometric identifiers were created for the respective scientific papers indexed in the Web of Science, the impact of these papers was analysed using altmetric indicators. A multiple linear regression model was developed to relate the variables from each information source in an attempt to detect the ones that contribute to the social impact of scientific production. The preliminary results show that the publishing journal, type of paper and international collaboration are the most significant factors in that regard.

Keywords
Altmetrics, research projects, scientific information, social networks

Conference Topic
Altmetrics

Introduction
Until only a few years ago, the predominant model for the production and assessment of academic knowledge revolved around the dissemination of research results in the form of published papers and patents. Against that backdrop, scientometrics and bibliometrics provided the tools of choice for analysing scientific productivity and studying the impact of research within the academic community (Callon et al., 1995). In addition, the analysis of research projects has proven to be a valuable approach to locate data on the lines of research proposed and accepted and acquire insight into significant aspects of research underway (Plaza, 2001).

While scientometrics and bibliometrics are consolidated disciplines for studying scientific output and impact, the traditional model of scientific dissemination has been revolutionised by the institution of the Web 2.0 and the furtherance of inter-individual communication and collaboration. Fora, blogs and social networks such as Facebook or Twitter that have proven to be so popular among the public at large have also influenced the members of the scientific community, giving rise to a so-called 'academic social network'. The proliferation of informal communication vehicles poses a new challenge for the analysis of scientific activity (Mohammadi and Thelwall, 2013) that calls for new approaches to the issue. One might consist in supplementing traditional scientometric studies with the formulation of altmetric indicators that constitute a measure of public interest in R&D+I in a particular area.

Given the existence of the various avenues for scientific communication, the present study tracked the flow of scientific knowledge, from the R&D projects themselves to the publication of the findings in scientific journals and the possible impact in social networks. A dual aim was pursued. The production of new knowledge was explored by analysing its
various stages, and a model was developed to identify the factors that contribute to the social impact of scientific projects and papers.

The specific field chosen for the analysis was ‘energy savings’, regarded as a key factor in countries’ economic development, with a heavy social and environmental impact that cuts across a host of industries. It is also a sector under constant development due to the intense technological innovation involved. One of the aspects that arouses greatest interest and that has been crucial for the development of national and international R&D+I policies such as implemented by the European Union is the relationships among energy efficiency, sustainability and the use of renewable resources. The OECD has also sought to heighten awareness of the importance of sustainable development in a number of programmes geared to improving study and promotion of the conceit, particularly in the area of sustainable and so-called ‘green’ growth (OECD, 2011; OECD, 2012).

Sources and methodology

Three sources of information were used in the study.

European projects-CORDIS: the calls for projects under the European Union’s Framework Programme have become one of the main avenues for scientific and technological activity and funding for the region’s institutions. FP7 introduced a change in approach to international cooperation, placing greater emphasis on R&D activities with third countries to rise to global challenges. The CORDIS database (http://cordis.europa.eu/projects/home_es.html) was used as a source of information on energy savings research channelled through FP7 projects. The following information was retrieved for each project: reference, start and finish date, abstract, lead institution, participating institutions, participating countries, total funding, funding obtained.

Papers-Web of Science: the analysis of information flows entailed studying the scientific papers stemming from the projects analysed. The database used was the international and multi-disciplinary Web of Science, which lists the scientific output deemed to be of highest quality and prestige. Information was retrieved from the Core Collection, which includes the SCI, SSCI and A&HCI databases. Papers on energy savings projects were identified by searching funding-related fields (FO=funding agency; FG=grant number; FT=funding text) for the projects analysed under those tags. The data collected were cleansed and the main activity indicators were found (number of papers, language, paper type, journal, specialisation (subject area), collaboration (number of authors, institutions and countries), impact (citations) and visibility (publishing journal quartile).

Social networks-ALTMETRICS.com: with the main bibliometric indicators in hand, altmetric indicators were identified through papers’ DOIs. A script developed by the Carlos III University of Madrid's Metric Information Studies Laboratory was deployed to obtain information from the ALTMETRICS.com platform on: posts, Twitter, Wikipedia, communication media, Google and Facebook and Mendeley, CiteUlike and Connotea readers.

The data collected in the initial stage furnished sufficiently descriptive information to characterise the results of each of the aforementioned sources.

The second stage consisted in inter-relating all the information obtained to develop a multiple linear regression model with which to determine links among the indicators analysed and detect possible relationships among certain project characteristics, the respective papers and social impact. To be able to run statistical tests, the model had to deliver a quantitative variable. Although the Altmetric.com ‘score’ could have been used, it was ruled out because its calculation is based on academic rather than social impact variables, such as the number of
academic social network (such as Mendeley or CiteULike) readers and the weights assigned to the variables are somewhat arbitrary.

Factor analysis techniques were consequently applied to obtain a single, more representative variable to quantify the social impact of research. Based on the contribution of each altmetric indicator to this new fictitious variable, a system of weights was established to formulate a synthetic indicator, following OECD (2008) recommendations, to be applied to scientific papers.

Defined as ‘score2’, this synthetic variable was used to build a multiple linear regression model, into which the variables and factors given in Table 1 were entered. In light of the colinearity between some of the possible variables (number of authors and institutions; number of countries and international collaboration), a selection was made to narrow redundancies and develop a sounder model.

Table 1. Variables used in the initial multiple linear regression model

<table>
<thead>
<tr>
<th>Item</th>
<th>Variable name</th>
<th>Type</th>
<th>Explanation</th>
<th>Possible values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Project_cost</td>
<td>Numerical</td>
<td>Total cost of the project</td>
<td>1:n</td>
</tr>
<tr>
<td>Paper</td>
<td>SO</td>
<td>Factor</td>
<td>Source of publication</td>
<td>Name of journal</td>
</tr>
<tr>
<td>Paper</td>
<td>DT</td>
<td>Factor</td>
<td>Document type</td>
<td>Article; Proceedings; Review</td>
</tr>
<tr>
<td>Paper</td>
<td>WC</td>
<td>Factor</td>
<td>Web of Science categories</td>
<td>List of WoS categories</td>
</tr>
<tr>
<td>Paper</td>
<td>Numaut</td>
<td>Numerical</td>
<td>Number of authors</td>
<td>1:n</td>
</tr>
<tr>
<td>Paper</td>
<td>Colab</td>
<td></td>
<td>International collaboration</td>
<td>Yes; No</td>
</tr>
<tr>
<td>Paper</td>
<td>Best quartile</td>
<td>Factor</td>
<td>Best quartile in JCR</td>
<td>Q1; Q2; Q3; Q4</td>
</tr>
</tbody>
</table>

The formula for the initial regression model was:

\[
\text{score2} \sim \text{project_cost} + \text{SO} + \text{DT} + \text{WC} + \text{numaut} + \text{Colab} + \text{best\_quartile}
\]

The preliminary results for the two stages are discussed in the section below.

Preliminary results

Stage 1

The descriptive information for the first stage is summarised in Figure 1.

The findings showed that scientific papers were generated by two-thirds of the projects analysed, to the exclusion of 88. Project papers were highly concentrated, with just 11 projects (4% of the total) accounting for 560 articles (33% of the total). Whilst the vast majority of the papers had a DOI, only 22% had any impact on social networks.

Figure 2 shows the projects grouped by subject area and the number of papers stemming from each (size of nodes), as well as the presence or absence of altmetric indicators.

Stage 2

Factor analysis yielded a first variable that explained 73% of the variability. Percentage weights were assigned to each altmetric indicator based on their contribution to this new fictitious variable (Table 2). The synthetic indicator denoting the social impact of papers was calculated on the grounds of those weights.
Figure 1. Projects analysis findings: papers and mentions in social networks

<table>
<thead>
<tr>
<th>Projects</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total FP7 projects (2007-2013)</td>
<td>25 690</td>
</tr>
<tr>
<td>No. of sub-programmes</td>
<td>23</td>
</tr>
<tr>
<td>No. of subject areas</td>
<td>67</td>
</tr>
<tr>
<td>No. of projects on energy savings</td>
<td>265</td>
</tr>
<tr>
<td>Energy savings-related sub-programmes</td>
<td></td>
</tr>
<tr>
<td>Energy (125); ICTs (80); Transport (21)</td>
<td></td>
</tr>
<tr>
<td>Countries with highest participation</td>
<td></td>
</tr>
<tr>
<td>Germany, France, England, Spain, Italy</td>
<td></td>
</tr>
<tr>
<td>Countries with highest leadership rate</td>
<td></td>
</tr>
<tr>
<td>Ireland, Spain, Germany, Italy, France, Austria</td>
<td></td>
</tr>
<tr>
<td>Average No. of institutions per project</td>
<td>12</td>
</tr>
<tr>
<td>Maximum No. of institutions per project</td>
<td>45</td>
</tr>
<tr>
<td>Funding range</td>
<td>Min: 261 451.80 / Max: 35 499 975.65</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Papers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total papers</td>
<td>1,654</td>
</tr>
<tr>
<td>Main WoS categories</td>
<td></td>
</tr>
<tr>
<td>Energy Fuels (23 %); Materials Science Multidisciplinary (21 %); Physics Applied (19 %); Chemistry Physical (13 %); Engineering Chemical (10 %)</td>
<td></td>
</tr>
<tr>
<td>Main publishing journals</td>
<td></td>
</tr>
<tr>
<td>Countries with highest publication</td>
<td></td>
</tr>
<tr>
<td>Germany (20 %); Italy (18 %); England (15 %); Spain (13 %); Netherlands (11 %); Switzerland (11 %); France (10 %)</td>
<td></td>
</tr>
<tr>
<td>Distribution by quartile (best quartile per paper)</td>
<td>Q1 (31 %); Q2 (20 %); Q3 (6 %); Q4 (3 %)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Social networks</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total papers with DOI</td>
<td>1,641</td>
</tr>
<tr>
<td>Documents with altmetric indicators</td>
<td>354 (21.57 %)</td>
</tr>
<tr>
<td>Most frequent altmetric indicators</td>
<td></td>
</tr>
<tr>
<td>Post (100 %); Mendeley (69 %); CiteULike (69 %); Tweets (68 %); Facebook (20 %)</td>
<td></td>
</tr>
</tbody>
</table>

![Diagram showing analysis of projects, papers, and social networks]
Table 2. Weight assigned to each altmetric indicator

<table>
<thead>
<tr>
<th>Altmetric indicator</th>
<th>Weight (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cited_by_fbwalls_count</td>
<td>14.6</td>
</tr>
<tr>
<td>cited_by_tweeters_count</td>
<td>16.2</td>
</tr>
<tr>
<td>cited_by_feeds_count</td>
<td>15.6</td>
</tr>
<tr>
<td>cited_by_rdts_count</td>
<td>4.3</td>
</tr>
<tr>
<td>cited_by_msm_count</td>
<td>16</td>
</tr>
<tr>
<td>cited_by_wikipedia_count</td>
<td>7.6</td>
</tr>
<tr>
<td>cited_by_videos_count</td>
<td>10.1</td>
</tr>
<tr>
<td>cited_by_gplus_count</td>
<td>15.6</td>
</tr>
</tbody>
</table>

As the analysis of variance of the initial model only yielded three variables with a p-value of under 0.05, journal (‘SO’), paper type (‘DT’) and international collaboration (‘Colab.’), these were the variables selected to build a simpler model, the ANOVA for which is given in Table 3.

Table 3. Analysis of variance and p-value for each variable

<table>
<thead>
<tr>
<th>Analysis of Variance Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Call: score2 ~ SO + DT + Colab</td>
</tr>
<tr>
<td>Response: score2</td>
</tr>
<tr>
<td>Df</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>Colab</td>
</tr>
<tr>
<td>DT</td>
</tr>
<tr>
<td>SO</td>
</tr>
<tr>
<td>Residuals</td>
</tr>
</tbody>
</table>

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

One prominent characteristic of the final model, its negative y-intercept (-3), inferred that the a priori likelihood of a paper generating social impact is very low. With international collaboration, however, the slope on the social impact prediction curve rose by a factor of +3, varying as well with paper type. Papers in the form of articles or proceedings were penalised by a factor of -5, whereas reviews were flattered by a factor of +13. The publishing journal was the most influential variable. The model showed that most journals (507) affected the slope positively, although a few (4) had an adverse effect and yet others (7) were neutral. On average, journals contributed with a value of 38.7 to the model’s standardised regression coefficient, beta. The ten journals with the highest beneficial effect were: Science; Earth System Science Data; Nature Genetics; Nature; Trends in Genetics; Genome Biology; Microbial Cell Factories; BMC Genomics; Nature Communications; Biotechnology for Biofuels.

In addition, due the score variable2 does not met the normality assumption, non parametric tests were calculated in order to get effect size and p-value of each variable (table 4). For ‘Colab’ variable, Mann-Whitney U test was used, because is a dicotomical variable (yes/no), but in the case of ‘DT’ and ‘SO’ variables were necessary to use Kruskal-Wallis test, and according to Tomczak & Tomczak (2014), $E^2$ was calculated so as to know the effect of these
variables, which does not show many differences with ANOVA test, due to the bigger effect size is in ‘SO’ variable and is low in ‘Colab’ and ‘DT’ variables.

Table 4. Addition test with effect size and p-value for each variable

<table>
<thead>
<tr>
<th>Variable</th>
<th>Test</th>
<th>Effect size</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colab</td>
<td>Mann-Whitney U</td>
<td>0.046</td>
<td>0.03018</td>
</tr>
<tr>
<td>DT</td>
<td>Kruskal-Wallis</td>
<td>0.0097 (E²)</td>
<td>0.001006</td>
</tr>
<tr>
<td>SO</td>
<td>Kruskal-Wallis</td>
<td>0.53 (E²)</td>
<td>&lt; 2.2e-16</td>
</tr>
</tbody>
</table>

Conclusions
In light of the importance of energy savings on the scientific as well as the social, economic and environmental scales, research is in order to determine how scientific and technological activity is being conducted and identify the aspects of greatest significance for academics and citizens. Scientometric techniques constitute a promising approach to this issue. This study identified the key characteristics of energy savings projects funded under the Seventh Framework Programme and analysed the flow of knowledge from project to publication in scientific journals and from there to social networks.

The preliminary findings show that papers have been published in WoS-listed journals around two-thirds of the projects. Although that figure may seem low, many of the projects underway in 2013-2014 are presently being written up. The high concentration of papers in a short number of projects is striking, however. Altmetric indicators show that the number of papers mentioned in social networks (22 %) is still low and apparently unrelated to productivity; i.e., the projects with the largest number of papers are not necessarily the ones cited in social networks.

The study’s major contribution is indisputably the development of a multiple linear regression model to detect relationships among variables that help understand the social impact of papers. The initial findings show that the likelihood of mention in social networks is low and favoured most by three factors: publishing journal, paper type and international collaboration. The next stage in the research underway will be to analyse these findings in greater depth.

Acknowledgements
This study was conducted under the project entitled ‘Detection of new research and innovation fronts in energy efficiency. Analysis of knowledge flows in the scientific domain, industry and society’ (ref.: CSO2014-51916-C2-1-R), funded by Spain’s Ministry of the Economy and Competition.

References
Assessing the Interdependencies between Scientific Disciplinary Profiles at the Country Level: a Pseudo-Likelihood Approach

Cinzia Daraio, Francesco Fabbri, Giulia Gavazzi, Maria Grazia Izzo, Luca Leuzzi, Giammarco Quaglia and Giancarlo Ruocco

Abstract
The investigation of the dynamics of national disciplinary profiles is at the forefront in quantitative investigations of science. There is an increasing number of papers that analyse the disciplinary specialization at the country level. We contribute to this literature by proposing a new approach to investigate the complex interactions among scientific disciplinary profiles. The approach is based on recent pseudo-likelihood techniques introduced in the framework of machine learning and complex systems. We infer, in a Bayesian framework, the network topology and the related interdependencies among national disciplinary profiles. We provide an illustration on data extracted from the Scopus database which relate to the national scientific production of most productive world countries for the 27 Scopus subject categories.

Conference Topic
Country-level studies, Methods and techniques

Introduction and Method
The dynamics of national research systems is a topical issue in quantitative science and technology research. The number of works on this issue has seen a considerable increase from the King (2004)’s work until the most recent years. Country-level studies of the evolution of disciplinary profiles include Glänzel (2000); Glänzel & Schlemmer (2007); Glänzel et al. (2006, 2008); Hu & Rousseau (2009); Tian et al. (2008); Wong (2013); Wong et al. (2012); Yang et al. (2012); Radosavic & Yoruk (2014); Bongioanni et al. (2014, 2015), Shen et al. (2016) and Li (2017). In particular, Shen et al. (2016) extend the Input-output model of the economist Leontief to investigate the interrelations between the scientific subfields of physics. We consider research systems as complex systems. Once the analogy is well defined the mathematical tools developed for complex systems can be exploited for studying research systems. Science is considered a complex system also according to the sociological perspective (see e.g. Shi et al. 2015) based on the actor network theory (Latour, 2005). Networks are general models, which can represent the relationships within or between given systems. The structure and function of complex networks is widely studied in the statistical mechanics: see e.g. the classical reviews by Albert and Barabasi (2002) and Newman (2003) and the recent book by Barabasi (2016). Network approaches are widely applied in scholarly evaluation. West and Vilhena (2014) provide a clear introduction to the topic, which has become a “cornerstone of bibliometric research” since the seminal work of Price (1965). The most studied networks include paper-level citation networks (in which nodes are the papers and the links are the citations between the papers) and co-authorship networks (in which nodes are the authors and the links represent the frequency a pair of authors has coauthored). West and Vilhena (2014, p.
conclude their overview stating that: “Network-based measures are more complicated than non-network measures, but the richness gained with such a measure is worth the extra effort.” The prediction of future links or the reconstruction of missing links from an incomplete network is a related interesting stream of literature (see Guns, 2014). In this paper we adopt a different level of analysis. In our modelling, nodes are the disciplinary profiles (or the scientific production in a given research area) of countries and the links represent the interdependencies existing between them. In addition, we consider these links as unknown parameters. We infer these unknown parameters by applying recently introduced tools to solve inverse problems in graphical models with wide applications to complex systems. The approach that we propose in the present paper is based on the similarities between statistical physics models for complex systems and research systems. See Table 1.

Table 1: Analogies between Physics of complex systems and research systems

<table>
<thead>
<tr>
<th>Concept in the framework of statistical physics</th>
<th>Concept in the framework of scientific system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node vectorial variable, “spin” s_i</td>
<td>Country’s disciplinary profile</td>
</tr>
<tr>
<td>Node variable components s_i = {s_i,γ}</td>
<td>Scientific disciplines in the disciplinary profile</td>
</tr>
<tr>
<td>Node interactions or couplings: J_{i,j}</td>
<td>Country-to-country disciplinary profile interdependencies</td>
</tr>
<tr>
<td>Hamiltonian (a function which represents the energy to be minimized): H</td>
<td>Generalized cost (social energy) function</td>
</tr>
<tr>
<td>Classical spin model with pairwise interaction</td>
<td>Pairwise country interdependencies</td>
</tr>
<tr>
<td>Generalized multicomponent spin model with arbitrary network, implemented in this paper: pairwise interactions.</td>
<td>Generalised interdependencies between country, discipline by discipline.</td>
</tr>
<tr>
<td>Hamiltonian:</td>
<td>Social energy or cost to minimize to determine the optimal solution.</td>
</tr>
<tr>
<td>[ H = \frac{1}{2} \beta \sum_{i,j} J_{i,j} s_i^\mu s_j^\mu - \sum_{i=1}^N s_i^\mu \cdot h_i ]</td>
<td></td>
</tr>
<tr>
<td>( \beta ) is the inverse of the temperature;</td>
<td>( \beta ) is an external control parameter</td>
</tr>
<tr>
<td>s_{i,γ} is the component γ of the vector variable of the node i.</td>
<td>s_i,γ denotes the discipline γ of country i;</td>
</tr>
<tr>
<td>i = 1, ..., N</td>
<td>s is a vector the elements of which contain the fluctuations around the world average of the number of publications in the various disciplines ranging from γ = 1, ..., D, where D = 27 Scopus subject categories</td>
</tr>
<tr>
<td>γ = 1, ..., D</td>
<td>N= 53 largest science producer countries</td>
</tr>
<tr>
<td>( \mu = 1, ..., M ) set of data</td>
<td>h_i denotes contextual variables of country i</td>
</tr>
<tr>
<td>h_i = external magnetic field on i.</td>
<td></td>
</tr>
<tr>
<td>Network of pairwise interacting spins (inferred by solving an inverse problem)</td>
<td>Network of interdependencies among disciplinary profiles (or scientific disciplines) at the country level</td>
</tr>
</tbody>
</table>
We exploit these analogies (Table 1) to model the interdependencies within the world research system as interactions. *Interaction* in physics is considered as a direct - and reciprocal - effect of one entity on one or more entities. The nature and the strength of this effect can be measured by applying tools developed by the statistical physics of complex systems. The concept of interaction in physics can find its correspondence in the one of *interdependency or interrelations* for research systems. The latter means the existence of a mutual *influence* among countries scientific activity: all countries are to some degree affected by the research activity of all other countries. This influence can be considered discipline by discipline or on the basis of the overall scientific production (disciplinary profile).

The term interrelation here encompasses all the channels of contact, exchange and so on, between two countries, which effect the convergence of the disciplinary profiles of the two countries. The interaction parameters of the *generalized multicomponent spin model* adopted here are thus effective parameters embedding several effects. Given the Hamiltonian related to this model (see Table 1), a positive interaction between two countries actually would lead to convergence of their disciplinary profiles in order to satisfy the principle of minimum energy. However, since the Hamiltonian represents a disordered many-body system with pairwise interactions, competition can arise between them giving rise to so-called *frustration*, that is a spin blocked between two opposite profiles, which is not able to choose the profile to follow grounding on the principle of minimum energy. The underlying hypothesis of our model is that a positive interaction would led to alignment between disciplinary profile, but not *vice versa*, that is the observation of alignment does not necessarily imply a positive interaction as well as the observation of misalignment does not exclude it. Our choice of the *generalized multicomponent spin model* is led by simplicity and because it guarantees the possibility to borrow the rigorous methodology developed by Boltzmann machine learning.

The aim of the present analysis is to derive the level of and the structure of these interactions. It is an inverse problem because the inference of the interactions is drawn from a set of data on the disciplinary profiles of countries. Judge and Mittelhammer (2011) clearly describe that inverse problems arise when one want to recover information on model parameters, i.e., coupling constants, by means of measurements of observable data. The solution of an inverse problem offers a connection between the data directly observed and the unknown information on model parameters. In an inverse problem, the model which generates the observed data is an input of the theory. The model can be gained by an a priori knowledge or hypothesized on general insights. In the latter case its strength can be a posteriori verified by testing its prediction.

We propose a new approach to make inference on the network topology and the related interdependencies between country-disciplinary profiles. The approach is developed in a Bayesian framework and relies on some recent pseudo-likelihood techniques introduced in the physics of complex systems (Ravikumar et al. 2010, Aurell & Ekeberg 2012, Tyagi et al. 2016; Marruzzo et al. 2016). In the following we briefly outline the proposed approach. For more details the reader is referred to the Appendix A. According to the Bayes theorem (see e.g. Barber, 2012):

$$p(\{J\}|\{s\}) = \frac{p(\{s\}|\{J\})p(\{J\})}{p(\{s\})} = \frac{p(\{s\}|\{J\})p(\{J\})}{\int_{\{J\}} p(\{s\}|\{J\})p(\{J\})}.$$  

where \(\{s\}\) are the data and \(\{J\}\) are unknown parameters of a given model, \(p(\ )\) states for a probability distribution and \(p( | \ )\) for a conditional probability distribution. We observe that \(\int_{\{J\}} p(\{s\}|\{J\})p(\{J\})\) is constant with respect to \(J\). The maximum of the conditional probability \(p(\{J\}|\{s\})\), is equal to the maximum of \(p(\{s\}|\{J\})\) the so called likelihood function, if one
assumes uniform the prior belief about the model parameters, \( p(J) \). We consider our system as a disordered system at equilibrium, described by a generalized multicomponent spin model. To this aim we define our variables as normalized shares of publications, which are the equivalent of the spin configurations, defined as:

\[
    s_i^{(\gamma)}(t) = \frac{\Delta_i^{(\gamma)}(t)}{\sqrt{\sum_{i=1}^{N} \Delta_i^{(\gamma)}(t)}}; \quad \Delta_i^{(\gamma)}(t) = n_i^{(\gamma)}(t) - \bar{n}^{(\gamma)}(t) = \frac{1}{N} \sum_{i=1}^{N} n_i^{(\gamma)}(t),
\]

where \( \gamma = 1, ..., D, t = 1, ..., T \), \( n_i^{\gamma} \) are the shares of articles published in a subject category \( \gamma \) for a given country \( i \), for \( i=1,...,N \), over the period \( t=1,...,T \) (here 1996-2012). They have the property that \( \bar{s}_i^{(\gamma)} = 0 \) and \( (\bar{s}_i^{(\gamma)})^2 = 1 \). In this way we account for the recent trend of increasing scientific production all over the world. By the normalization reported above in fact we define as variables only instantaneous fluctuations around the world average production in each given discipline at one data sample recording. Though the average scientific production is not constant in time the distribution of the deviations around the means can be considered as such. The assumption of equilibrium underlies a Boltzmann-Gibbs distribution,

\[
p(J, s) = \frac{1}{Z(J)} e^{-H(s|J)},
\]

where \( H \) is the Hamiltonian of the generalized Ising model (defined in Tab. 1) and \( Z \) is the partition function (normalization factor). It is also possible to draw a link between equilibrium and symmetry of the pairwise interactions. Symmetric couplings lead to a steady state described by the Boltzmann distribution while asymmetric ones to a non-equilibrium state (Krapivsky, 2010). It is also possible to assign to the system a particular dynamics, which leads it to a given steady state distribution. Recent developments achieved for dynamical inverse Ising model (Decelle, 2016, Nguyen, 2017) could represent an interesting extension of the present work, which is left for future research.

The Hamiltonian is obtained by a generalization of the Ising model, originally introduced to describe the behavior of ferromagnetic systems. In its more general formalism the Ising model can also account for a (parameter independent) weight, \( \beta \), and external biases, \( h_i \). When referred to magnetic systems, \( \beta \) is the inverse of temperature and \( h_i \) the magnetic external field. For sake of simplicity we fix here \( \beta = 1 \) and \( h_i = 0 \) for \( i \in (1, N) \). The Ising model has been largely applied in different fields, such as modelling the behaviour of magnets in statistical physics (Brush 1967), image processing and spatial statistics (Besag 1986, Geman 1984, Greig 1989), modelling of social networks (Banerjee 2008). Our choice is thus mainly justified from the possibility to recover and generalize tools developed and already tested in such different contexts. The Ising model, furthermore, is a classical example of a graphical model in exponential form, to which the Boltzmann machine learning (or approximation method related to it) can be applied, as discussed in Appendix A.

To solve the inverse problem we have to maximize the log-likelihood with respect to the set of parameters \( J \):

\[
l(J) = \log(L(J)) = \sum_{t=1}^{T} -H(s_t|J) - T \log(Z(J)).
\]

This is computationally hard to solve and for this reason a pseudolikelihood approach, which involves Boltzmann machine learning, is applied. The solution of the inverse problem consists in finding the optimal values of the set of parameters \( J \), which are supposed to generate the observable set of data. We can obtain the values of the set parameters \( J \) representing the interrelations between a pair of countries i) discipline by discipline by maximizing the discipline-dependent Log-Likelihood function, ii) for the full disciplinary profile by maximizing the Log-Likelihood function related to a cost function, which in turn depends on
the scalar product between the vectorial variables whose generic element refers to a given discipline (see Tab. 1 and Appendix A for details).

Given a set of parameters \{J\}, a zero value of the parameter \(J_{ij}\) between the pair \(ij\) means that the two countries are not interacting, a positive value indicate a tendency to align towards the same disciplinary profile, and a negative value, instead, shows a tendency towards ‘opposite’ disciplinary profile.

The following sections describe the data available and the main results of the analysis. The final section concludes the paper and outlines directions for further research.

Data

Data was extracted from the Scopus database and relate to the scientific production of world countries for 27 Scopus subject categories from 1996 to 2012.

Data problems in bibliometric studies are well known. A common way to reduce them is to analyse macro-level bibliometric data. According to Nederhof (1988) comparative analysis are more reliable when the unit of analysis is more aggregated because in a larger sample size, micro random errors mutually compensate.

Another issue of concern is given by the changes in coverage from inclusion or exclusion of journals which affect more small countries with less publications. This may lead to unreliable values when a country only has a small number of scholarly outputs (see e.g. Schubert et al. 1989). To avoid this problem, we investigate the disciplinary profiles of the 53 most productive countries, which account for more than 95% of the world scientific production in the considered period. To increase the number of available data we transformed yearly data into weekly data by means of a linear interpolation. The final number of observations considered is 833 and refers to weekly production of the number of articles (\(P\)) over the period 1996-2012. See Figure 1 for an illustration.

![Figure 1](image)

**Figure 1** Trends of the interpolated scientific production (\(P\)) in three disciplines: Physics and Astronomy, Computer Science and Medicine. For each discipline, the eight most productive countries are reported.

Results

Figure 2 shows the estimated interactions obtained by applying the method described in the previous section. We show only interactions strong enough, which significantly contribute to the cost function \(H\) (see Fig. 2). We infer a trend to clustering among countries belonging to a given geo-political or cultural area. By observing the interactions obtained for the network model with the overall Hamiltonian (i.e. the one describing the interrelations among disciplinary profiles) shown in Fig. 2 Panel IV we report a strong positive interaction between USA, Great Britain and Korea, whereas the USA interactions with all the other countries are
negative. Great Britain, in turn, weakly interacts with the other European countries, showing a positive interaction only with France and (quite weakly) with Spain. The Central Europe countries positively interact among them. When related to the recent historical context, it could result significant also that both Ukraine and Poland get a positive interaction with Russia. Taiwan, Korea and Singapore share positive interactions among each others.
Brazil, against trend, gets positive interactions with several countries world-spread but with none of the others Latin America countries. These results are in qualitative agreement with a quantitative measure of the overlap between disciplinary profiles of research systems discussed in Bongioanni et al. (2014). There, for example, a globalization of science in Europe was observed with a departure from the ‘European model’ by the leading countries such as Great Britain and Netherlands. By moving to the discipline-dependent interactions, i.e. the ones obtained by considering a discipline-dependent Hamiltonian, shown in Fig. 2 Panels I-III for representative scientific disciplines, it appears a larger number of positive interactions, in particular of USA with all the other world-spread countries. This could confirm a trend of globalization of science affecting not only Europe but all the world. We observe, e.g., that the interaction of USA with Russia, Japan and several European countries is in this case positive. The interaction of China with several countries is also positive, whereas for the interactions related to the overall Hamiltonian most of them are negative. In the case of Medicine, China disappears from the network, apart for a positive interaction with USA, probably due to a reduction of the number of articles published in this field. Figure 3 shows some examples of the outcome of our approach, which is the inferred network topology of selected disciplines.

![Graphical representation of the network topology](image)

**Figure 3.** Graphical representation of the network topology, for selected disciplines (Physics and Astronomy, Computer Science and Medicine), inferred from the values of the $J_{ij}$ parameters obtained by the maximization of the pseudo-likelihood function. The graphs report the most relevant interactions, i.e. those with $|J_{ij}| > 0.1 \max(|J_{ij}|)$. The diameter of the node representing the i-th country is proportional to the number of interactions $J_{ii}$. The thickness of the edge depends on the intensity of the related interaction (thicker edge means larger intensity).

Discussion and conclusion

The results presented in the previous section show the potential of the approach we propose. The policy of scientific research is an important subject. The requirements and the disciplinary orientation of research funding programs are incentive to creativity as well as to the scientific production of scientists (Azoulay et al. 2011). Recent studies on the economics and organization of science have highlighted the lack of tools and analysis to determine how to allocate the funds among different disciplines (Antonelli et al. 2011). Of course, it is the policy of research that has to choose how to distribute resources across different scientific areas. The approach we propose in this paper may provide a useful tool to support policy decisions. Once the reliability of the proposed approach is established, it is possible to exploit the additional features of the generalized multicomponent spin model, e.g. to include in the analysis some measures of country contextual variables through the external field $h_i$. For instance, an interesting extension of the study (left for future work) may be the inclusion of an external field given by a vector of research funding with elements country- and discipline-
dependent, which together with the mutual interactions between countries can direct the
disciplinary profile in a direction or in another.
Once the inverse problem is solved, the obtained results could be used to simulate the impact
of different levels of the funding (through the external field) on the disciplinary profiles
emerging as outputs of the generalized multicomponent spin model which considers the
interactions fixed to the one inferred, and change the external funding vector. This could
guarantee a larger and more solid predictive capability than the simple extrapolation of observed trends. Other interesting developments of this paper, which are left to future research, include:
-maximization of the Pseudo-Likelihood function with different samplings, e.g. accounting for country’s size;
-application of the approach to other indicators of citations, highly cited publications and so on to compare the estimated interdependencies and the network topologies obtained;
-refinements of the decimation (Marruzzo et al., 2016) to infer the network structure;
-overcome the limitations of the interpolation of data (to increase the number of observations) analyzing alternative data, such as downloads, which were not available for the present study;
and finally considering the intensity of the auto-interdependence of countries (through the J_ii), the so called “chemical potentials” of countries.

Acknowledgements

This work was supported by Elsevier that provided the data within the Elsevier Bibliometric Research Programme (EBRP) and by the project Sapienza Awards no. 6H15XNFS.

References

Bongioanni I., Daraio C., Ruocco G. (2014), A quantitative measure to compare the disciplinary profiles of research systems and their evolution over time, Journal of Informetrics, 8 (3), 710–727.
I. APPENDIX A

In this appendix we discuss the methodology used to obtain the parameters of the maximum Log-Likelihood function introduced in the paper. Firstly, we discuss the general grounds of the validity of the method used. Secondly, we deal with the application to our specific case.

Given the set of data, $\{X_n, n = 1, 2, ..., N\}$, assuming that the data are independent, and once defined the generative model, the Log-Likelihood function, $l(\{J\})$ becomes

$$l(\{J\}) = \log(L(\{J\})) = \sum_{\mu=1}^{M} -H(\{s^\mu\}|\{J\}) - M \log(Z(\{J\})), \quad (A1)$$

where $\mu = 1, ..., M$ is the label for a set of $M$ data. The inference problem consists in determining the set parameters $\{J\}$ which maximizes the function in Eq. A1.

We consider here the expression of the cost function (Hamiltonian) for a multicomponent variable $s = (s_1, s_2, ..., s_J)$, $H(\{s\}|\{J\})$, given by

$$H(\{s\}|\{J\}) = -\frac{1}{2} \sum_{i,j} J_{ij} s_i \cdot s_j, \quad (A2)$$

with $J_{ij} = J_{ji}$. The symbol “$\cdot$” in Eq. A2 states for a scalar product. The presence of a scalar product ensures that orthogonal or quasi-orthogonal vectors (i.e. countries which have a number of publications whatever large but in different fields) will have a small weight in the cost function. The sum is extended to all couples of nodes $(i, j)$ with $i \neq j$. The partition function $Z(\{J\})$ is

$$Z(\{J\}) = \sum_{\{s\}} e^{-H(\{s\}|\{J\})}, \quad (A3)$$

the sum is extended to all possible configurations in the phase space of the set of variables $\{s\}$.

The calculation of the above partition function is too demanding from a computational point of view already for a small number of variables. For this reason, we resort to the pseudo-likehood approximation (Aurell 2012, Tyagi 2016, Marruzzo 2016). It consists in maximizing a Pseudo-Log-Likelihood function based on the local conditional Log-Likelihood function at each node (see Eq.A6) in place of the Log-Likelihood function. It is possible to show that the estimation of the parameters obtained by a Pseudo-Log-Likelihood maximization is consistent with the maximization of the Log-Likelihood function, that the two functions are maximized by the same set of parameters. The hypothesis under which this statement holds, i.e. the strict concavity of the Pseudo-Log-Likelihood function with respect to the k elements of the set of parameters, is not too strict (see Hyvarinen 2006). Furthermore it is possible to show that under such a hypothesis the Pseudo-Log-Likelihood maximization is exact (i.e. equivalent to the Log-Likelihood maximization) in the case of infinite sampling (Aurell 2012). An important advantage of the Pseudo-Log-Likelihood function is that it is possible to maximize it in polynomial time.

We consider then the likelihood built on the local conditional probability on each variable $i$, one by one, rather than Eq. (A1). The cost function (Eq. A2), is rewritten as

$$H(\{s\}|\{J\}) = -s_i \cdot \left[ \sum_{i,j} J_{ij} s_j \right] - \frac{1}{2} \sum_{k, j \neq i} s_k \cdot J_{kj} s_j$$

$$= -s_i \cdot A_i(\{J\}) - \sum_{k \neq i} s_k \cdot B_{ik}(\{J\}) \equiv H_i(\{s_i\}|\{s_\backslash i\}, \{J\}) + H_{\backslash i}(\{s_{\backslash i}\}|\{J\}), \quad (A4)$$

where $s_\backslash i$ indicates the set of all input-variables except the $i$-th. The functions $A_i(\{J\}) = \frac{1}{2} \sum_{j} J_{ij} s_j$ and $B_{ik}(\{J\}) = \frac{1}{2} \sum_{j \neq k} J_{ij} s_j$ have been introduced in Eq. A4. The cost functions $H_i(\{s_i\}|\{s_\backslash i\}, \{J\})$ and $H_{\backslash i}(\{s_{\backslash i}\}|\{J\})$ are implicitly defined. Analogously we can rewrite the partition function as

$$Z(\{J\}) = \sum_{\{s\}} e^{-H(\{s\}|\{J\})} = \sum_{\{s_i\}} e^{-H_i(\{s_i\}|\{J\})} \sum_{\{s_{\backslash i}\}} e^{-H_{\backslash i}(\{s_{\backslash i}\}|\{J\})} \sum_{\{s_{\backslash i}\}} e^{-H_{\backslash i}(\{s_{\backslash i}\}|\{J\})}$$

$$= \sum_{\{s_i\}} e^{-H_i(\{s_i\}|\{J\})} \sum_{\{s_{\backslash i}\}} e^{-s_{\backslash i} \cdot A_{\backslash i}(\{J\})} \sum_{\{s_{\backslash i}\}} e^{-H_{\backslash i}(\{s_{\backslash i}\}|\{J\})} \prod_{\gamma=1}^{D} \frac{\sinh(A_{\gamma}(\{J\}))}{A_{\gamma}(\{J\})}$$

$$= \sum_{\{s_i\}} e^{-H_i(\{s_i\}|\{J\})} Z_i(\{s_{\backslash i}\}, \{J\}) \quad (A5)$$
where the local conditional probability at the \( i \)-th node is
\[
p(s_i|\{s_{\setminus i}\}, \{J\}) = \frac{1}{Z_i(\{J\})} e^{-H_i(s_i|\{s_{\setminus i}\}, \{J\})},
\] (A6)

and the local partition function is \( Z_i(\{J\}) = \sum_{\{s_i\}} e^{-H_i(s_i|\{s_{\setminus i}\}, \{J\})} \).

Then, the weighted average over the ensemble of \( \{s_{\setminus i}\} \) configurations is substituted by the uniform average over the observed data sets \( \mu = 1, \ldots, M \). By defining \( \ell_i(\{s_i|\{s_{\setminus i}\}\}, \{J\}) = \log[p(s_i|\{s_{\setminus i}\}, \{J\})] \), the Pseudo-Log-Likelihood function is
\[
\lambda(\{J\}) = \sum_{\mu=1}^{M} \sum_{i=1}^{N} \ell_i^\mu(\{s_i^\mu|\{s_{\setminus i}^\mu\}\}, \{J\}) = \sum_{i=1}^{N} \ell_i.
\] (A7)

The gradient of the Pseudo-Log-Likelihood function with respect to the parameter \( J_{ij} \) is given by
\[
\frac{\partial}{\partial J_{ij}} \lambda(\{J\}) = \sum_{\mu=1}^{M} \left[ \frac{1}{2} \sum_{\mu=1}^{M} \frac{s_i^\mu \cdot s_j^\mu}{Z_i^\mu(\{J\})} - \frac{1}{Z_i^\mu(\{J\})} \frac{\partial}{\partial J_{ij}} Z_i^\mu(\{J\}) \right] = \frac{1}{2} \sum_{\mu=1}^{M} \left[ \frac{s_i^\mu \cdot s_j^\mu}{Z_i^\mu(\{J\})} \frac{\sum_{s_{\setminus i}} s_i \cdot s_j e^{-H_i(s_i|\{s_{\setminus i}\}, \{J\})}}{\sum_{s_{\setminus i}} e^{-H_i(s_i|\{s_{\setminus i}\}, \{J\})}} \right]
\] (A8)

\[
= \frac{1}{2} \left[ \frac{1}{M} \sum_{\mu=1}^{M} s_i^\mu \cdot s_j^\mu - s_i \cdot s_j >_{i,J} \right] = \frac{1}{2} [C_{ij} - \frac{1}{M} \sum_{\mu=1}^{M} s_i^\mu \cdot s_j^\mu >_{i,J}],
\]

where “ \( >_{i,J} \)” states for ensemble average calculated for the probability distribution \( p(s_i|\{s_{\setminus i}\}, \{J\}) \). It is possible to rephrase the term \( \frac{\partial}{\partial J_{ij}} \lambda(\{J\}) \) in the gradient of the Log-Likelihood function, obtaining
\[
\frac{1}{Z(\{J\})} \frac{\partial}{\partial J_{ij}} Z(\{J\}) = \frac{1}{Z(\{J\})} \sum_{\{s_i\}} e^{-H_i(s_i|\{s_{\setminus i}\}, \{J\})} \sum_{\{s_{\setminus i}\}} s_i \cdot s_j e^{-H_i(s_i|\{s_{\setminus i}\}, \{J\})} = \frac{1}{Z(\{J\})} \sum_{\{s_i\}} e^{-H_i(s_i|\{s_{\setminus i}\}, \{J\})} \sum_{\{s_{\setminus i}\}} s_i \cdot s_j e^{-H_i(s_i|\{s_{\setminus i}\}, \{J\})} >_{i,J} >_{i,J} >_{i,J}
\] (A9)

and
\[
\frac{\partial}{\partial J_{ij}} \ell(\{J\}) = \frac{1}{2} [C_{ij} - s_i \cdot s_j >_{i,J} >_{i,J}].
\] (A10)

By comparing Eq. A8 and A10 it is possible to infer that in the limit \( M \to \infty \) i) both the gradients go to zero for the set of parameters \( \{J\} \) generating the observed data, ii) \( \frac{\partial}{\partial J_{ij}} \lambda(\{J\}) \to \frac{\partial}{\partial J_{ij}} \ell(\{J\}) \). This finally establishes the consistency of the maximum Pseudo-Log-Likelihood estimator and, furthermore, its coincidence with the maximum Log-Likelihood estimator in the limit \( M \to \infty \).

The gradient of the Log-Likelihood function can be calculated exactly, thus facilitating the computational solution of the inference problem. The explicit expression of \( \frac{\partial}{\partial J_{ij}} \lambda(\{J\}) \) is reported in the following.

To deal with a lower number of parameters in place of maximizing the Pseudo-Log-Likelihood function, given by the sum of the single-node Pseudo-Log-Likelihood functions (Eq. A7) we maximize each single-node Pseudo-Log-Likelihood function. Since the couplings in the Ising model should be symmetric the final estimate of the \( J_{ij} \) parameter is obtained by taking the average \( (J_{ij} + J_{ji})/2 \). When a standard Pseudo-Log-Likelihood maximization is used, some couplings can be largely overestimated. To avoid such drawbacks we used a \( b_2 \) regularizer (Ravikumar 2010), i.e. in place of maximizing the \( \lambda(\{J\}) \) function we maximize the function \( \lambda(\{J\}) - b_2 (\sum_{i,j} J_{ij}^2)^{1/2} \), where \( b_2 \) is a suitable chosen constant.

The maximization of the single-node Pseudo-Log-Likelihood functions has been performed by means of the MATLAB fminunc package by selecting a trust-region optimization algorithm.

In the following we first rephrase the expression of the Log-Likelihood function by isolating the contribution of the \( i \)-th node. We finally calculate the gradient of the Pseudo-Log-Likelihood function with respect to \( J_{ij} \).

The sum \( \sum_{\{s_i\}} e^{-s_i A_i(\{J\})} \) in Eq. A5 has been calculated by assuming that the values of the \( i \)-th input variable can continuously vary in the interval \([-1, 1] \), obtaining
\[
Z_i(\{J\}) = \sum_{\{s_i\}} e^{-s_i A_i(\{J\})} = \int_{-1}^{1} ds_i e^{-s_i A_i(\{J\})} = \prod_{\gamma=1}^{D} \int_{-1}^{1} ds_{\gamma} e^{-s_{\gamma} A_{\gamma}(\{J\})} = \prod_{\gamma=1}^{D} \frac{2 \sinh(A_{\gamma}(\{J\}))}{A_{\gamma}(\{J\})}.
\] (A11)
Similarly, it is possible to write the function \( Z_{\gamma}(\{J\}) = \sum_{\{s_{\gamma}\}} e^{-H_{\gamma}(\{s_{\gamma}\},\{J\})} \), by exploiting the function \( B_{i,k}(\{J\}) \) defined above, obtaining
\[
Z_{\gamma}(\{J\}) = \sum_{\{s_{\gamma}\}} \prod_{i} e^{-H_{\gamma,i}(\{s_{\gamma}\},\{J\})} = \sum_{\{s_{\gamma}\}} \prod_{i} e^{-H_{\gamma,i}(\{s_{\gamma}\},\{J\})} \sum_{\{s_i,k\}} B_{i,k}(\{J\}) = \sum_{\{s_{\gamma}\}} \prod_{i} e^{-H_{\gamma,i}(\{s_{\gamma}\},\{J\})} \prod_{\gamma=1}^{D} \frac{\sinh(B_{i,k}^{(\gamma)}(\{J\}))}{B_{i,k}^{(\gamma)}(\{J\})}.
\]
(A12)

By iterating this procedure to the remaining variables it is finally possible to write the partition function as the product
\[
Z(\{J\}) = \prod_{\gamma=1}^{D} \frac{\sinh(A_{i}^{\gamma}(\{J\}))}{A_{i}^{\gamma}(\{J\})} \prod_{\gamma=1}^{D} \frac{\sinh(B_{i,k}^{(\gamma)}(\{J\}))}{B_{i,k}^{(\gamma)}(\{J\})} \frac{\sinh(F_{i,k}^{\gamma}(\{J\}))}{F_{i,k}^{\gamma}(\{J\})}.
\]
(A13)

The Log-Likelihood function becomes
\[
I(\{J\}) = \sum_{\mu=1}^{M} \sum_{i=1}^{N} [s_{i}^{\mu} \cdot A_{i}(\{J\}) + \sum_{k \neq i} s_{k} \cdot B_{i,k}^{(\mu)}(\{J\})] - M \sum_{\gamma=1}^{D} \left[ \log\left( \frac{\sinh(A_{i}^{\gamma}(\{J\}))}{A_{i}^{\gamma}(\{J\})} \right) + \log\left( \frac{\sinh(B_{i,k}^{(\gamma)}(\{J\}))}{B_{i,k}^{(\gamma)}(\{J\})} \right) \right] + \log\left( \frac{\sinh(F_{i,k}^{\gamma}(\{J\}))}{F_{i,k}^{\gamma}(\{J\})} \right),
\]
(A14)

The Pseudo-Log-Likelihood function, defined in Eqs. A6 and A7, takes now the expression
\[
\lambda(\{J\}) = \sum_{\mu=1}^{M} \sum_{i=1}^{N} [s_{i}^{\mu} \cdot A_{i}(\{J\})] - \sum_{\gamma=1}^{D} \log\left( \frac{\sinh(A_{i}^{\gamma}(\{J\}))}{A_{i}^{\gamma}(\{J\})} \right)
\]
(A15)

The difference between the Log-Likelihood function and the Pseudo-Log-Likelihood function clearly appears by comparing Eqs. A14 and A15.

We can now explicitly calculate the gradient of the Pseudo-Log-Likelihood with respect to the set of parameters \( J_{ij} \). From Eq. A8, we need to calculate the quantity \( \langle s_{i}^{\mu} \cdot s_{j}^{\nu} \rangle_{i,j}(\{J\}) \). It is (for sake of clarity the index \( \mu \) is omitted)
\[
\langle s_{i}^{\mu} \cdot s_{j}^{\nu} \rangle_{i,j}(\{J\}) = \frac{\sum_{\{s_{\gamma}\}} s_{i}^{\mu} \cdot s_{j}^{\nu} e^{-\sum_{\gamma} H_{\gamma}(\{s_{\gamma}\},\{J\})}}{\sum_{\{s_{\gamma}\}} e^{-\sum_{\gamma} H_{\gamma}(\{s_{\gamma}\},\{J\})}} = \frac{1}{Z(\{J\})} \sum_{\{s_{\gamma}\}} s_{i}^{\mu} \cdot s_{j}^{\nu} \int_{-1}^{1} ds_{i} \cdot s_{i} e^{-\sum_{\gamma} \sum_{\alpha=1}^{D} s_{i}^{\gamma} A_{i}^{\gamma}} = \frac{1}{Z(\{J\})} \sum_{\gamma=1}^{D} s_{i}^{\gamma} \int_{-1}^{1} ds_{i} \cdot s_{i}^{\gamma} e^{-\sum_{\gamma} \sum_{\alpha=1}^{D} s_{i}^{\gamma} A_{i}^{\gamma}} = \frac{1}{Z(\{J\})} \sum_{\gamma=1}^{D} s_{i}^{\gamma} \left[ \prod_{\gamma=1}^{D} \frac{2}{A_{i}^{\gamma}} \sinh(A_{i}^{\gamma}) \right] = \frac{2}{A_{i}^{\gamma}} (A_{i}^{\gamma} \cosh(A_{i}^{\gamma}) - \sinh(A_{i}^{\gamma})).
\]
(A16)

The expression of \( Z(\{J\}) \) is reported in Eq. A11. It is possible to rewrite it, for a given index \( \gamma \), as \( Z_{i} = 2 \sinh(A_{i}^{\gamma}) \prod_{\alpha \neq \gamma}^{1} A_{i}^{\alpha} \). By inserting this latter expression in Eq. A16, we obtain
\[
\langle s_{i}^{\mu} \cdot s_{j}^{\nu} \rangle_{i,j}(\{J\}) = \sum_{\gamma=1}^{D} \frac{s_{j}^{\gamma} (2 \sinh(A_{i}^{\gamma}) \prod_{\alpha \neq \gamma}^{1} A_{i}^{\alpha})}{2 (A_{i}^{\gamma})} (A_{i}^{\gamma} \cosh(A_{i}^{\gamma}) - \sinh(A_{i}^{\gamma})) = \left[ \frac{1}{\tanh(A_{i}^{\gamma})} - \frac{1}{A_{i}^{\gamma}} \right],
\]
(A17)

and finally
\[
\frac{\partial}{\partial J_{ij}} \lambda(\{J\}) = \frac{1}{2} M \left[ C_{ij} - \frac{1}{M} \sum_{\mu=1}^{M} \sum_{\gamma=1}^{D} s_{j}^{\gamma(\mu)} \left[ \frac{1}{\tanh(A_{i}^{\gamma(\mu)})} - \frac{1}{A_{i}^{\gamma(\mu)}} \right] \right]
\]
(A18)
Multidimensional assessment of R&D performance: evidence from the pilot evaluation exercise of Russian public research institutions

Konstantin Fursov ¹  Gleb Kuzmin ²

¹ksfursov@hse.ru
National Research University Higher School of Economics (Russia)

²gkuzmin@hse.ru
National Research University Higher School of Economics (Russia)

Abstract

Research evaluation recently became a widely disseminated exercise aimed in the end of the day at improving the cost efficiency of public funding of national R&D sectors. In November 2013, the Government of the Russian Federation initiated a national evaluation exercise of public research institutions (PRIs) to provide information basis for development of S&T policies aimed at increasing effectiveness and strengthening the role of R&D performing institutions in economic and social development. The aim of this paper is that of providing an approach for multidimensional assessment of R&D performance based on quantitative data derived from the national evaluation exercise, specifically looking at its applicability and limitations for further analysis and preliminary differentiation of PRIs as well as for use in policymaking.

Conference Topic
Science policy and research assessment

Introduction

Research evaluation of public institutes is considered important for measuring the performance of R&D sector, first of all, in terms of improving the cost efficiency of public funding. Various studies (e.g. see Luwel et al., 1999; Senker 2001; Coccia, 2004; Abramo & Angelo, 2015; Ancaiani et al., 2015) show a growing interest in evaluating national scientific performance with the use of different metrics and models.

In Russia the first round of the national evaluation exercise that covered around 400 PRIs subordinated to the national science academies and federal agencies was implemented in 2009-2011. This initiative involved a wide range of indicators (R&D expenditure, personnel, research equipment, publications, citations, journal impact-factors, patents and other IPRs, cooperation, innovation infrastructure, income and profitability, etc.), but its results did not lead to any substantial change neither in policy instruments nor in the institutional structure of the national R&D sector. As for the use of performance-based indicators for evaluating individual researchers, there was no common framework established at the national level, but a number of research institutions and universities adopted certain practices within their human resource development strategies.

In 2013, a new round of research evaluations was launched taking into account the lessons learnt from the previous round. The distinctive features of this latest initiative included (1) an open and interagency approach in selecting criteria for evaluation; (2) combination of quantitative and qualitative (peer-review based) assessment procedures applicable, if necessary, also at the level of research teams; (3) grouping of R&D institutions into “reference groups” on the basis of respective research fields and basic functions; (4) a 5-year
evaluation period. A list of indicators was revisited and comprised both conventional inputs (R&D expenditure and personnel) as well as research equipment and novel output categories. In particular, there were added different types of publications (journal articles, conference proceedings, books and book chapters), citations and impact-factors; results of inventive work like designs, blueprints, patents and other IPRs; and financial results such as income from technology transfer or S&T services.

While this exercise has not yet been finished and its results have not been implemented for decision-making, it allowed constructing an advanced model of knowledge production function taking into account different types of R&D outputs. Furthermore, the exercise stimulated ‘Agora processes’ (Barré, 2001, 2005) in the scientific community around the use of S&T indicators for research evaluation and motivated academics to accept expert roles in forthcoming peer-review procedures.

The aim of this paper is that of providing an approach for multidimensional assessment of R&D performance based on quantitative data, specifically looking at its applicability and limitations for further analysis and preliminary differentiation of organizations as well as for use in policymaking.

**Data sources and methods**

Data for this study is derived from the Federal System for Monitoring of R&D Performance maintained and updated by the Ministry of Education and Science of the Russian Federation (http://www.sciencemon.ru/) in September 2016. Reference year is 2015. Primary information was collected from institutions performing R&D in line with the guidelines explaining definitions of key terms and methodology for reporting key variables. The dataset available included information on different characteristics of R&D performance for the 1625 institutions including PRIs, universities and research divisions of state companies. For further analysis to provide higher homogeneity of objects under assessment only PRIs were selected. Additionally arithmetic and logical controls were conducted for key variables to construct a final sample covering 815 observations. Research outputs considered for further analysis were journal articles, conference proceedings and book chapters indexed in Web of Science; blueprints and designs, registered intellectual property rights (IPRs) and others. Additionally a number of financial results such as income from technology transfer, S&T services and contractual works provided were taken into account.

Preliminary correlation analysis demonstrated strong significant cohesion between basic research and publication output. At the same time, as seen from the table 1, organizations implementing applied research and experimental development are likely to be the main providers of S&T services and have weak linkages to IPR protection. In other words organizations with identical balance between different types of research activities may have different results and vice versa. Therefore, they cannot be included into a single and homogenous reference group and compared. Similarly, it could be shown that other characteristics such as type of institution, legal status of an organization or its size are weak to differentiate R&D outputs.

### Table 1. Pearson correlation matrix

<table>
<thead>
<tr>
<th>Variable</th>
<th>Publications</th>
<th></th>
<th>IPRs</th>
<th></th>
<th>S&amp;T services</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R&amp;D expenditure</td>
<td>abs*</td>
<td>norm*</td>
<td>abs</td>
<td>norm</td>
<td>abs</td>
<td>norm</td>
</tr>
<tr>
<td>Basic research</td>
<td>0.76**</td>
<td>0.06</td>
<td>0.14**</td>
<td>-0.20**</td>
<td>0.64**</td>
<td>-0.19**</td>
</tr>
<tr>
<td>% of total</td>
<td>0.10**</td>
<td>0.39**</td>
<td>0.61**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applied research</td>
<td>-0.13**</td>
<td>0.15**</td>
<td>0.15**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% of total</td>
<td>0.18**</td>
<td>0.23**</td>
<td>0.59**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In order to construct the reference groups for PRIs a two-step algorithm was implemented. At the first step, research organizations were divided into groups by field of science (the first level of the OECD FOS classification (OECD, 2002) in line with the specific research areas mentioned by the organizations. As a result, an organization may refer to more than one field of science. At the second step, within each field of science, a research profile of an organization describing its orientation towards one or several social functions, i.e. knowledge generation, technology development or provision of S&T services, and consequently corresponding with certain types of research outputs was identified. Herewith a research profile was determined with the use of the following key indicators:

A. The number of papers indexed in Web of Science per 100 researchers.

B. Number of IPRs registered in the Russian Federation or abroad as well as the number of issued design documentation per 100 researchers.

C. Income from contractual R&D, S&T services provided and technology transfer per total R&D personnel of an organization.

As seen from the above, to reduce scale effects each indicator (A – C) reflecting certain type of outputs is divided by the number of employees, mainly involved in its acquisition. While researchers tend to be mostly engaged in publications and IPR production, financial results require involvement of different categories of R&D personnel.

Consequently, an organization refers to a particular research profile reflecting its bent for certain research function and corresponding output, if one or more of the indicators mentioned above (A – C) is not zero and equal or exceeds the median value for the field of science. It allows identification up to eight research profiles (see figure 1). An intersection of field of science and a research profile constitutes a reference group.

![Figure 1. Research profiles of organisations](image-url)
demonstrating lower level of research productivity by all indicators. At the same time, it may have some other characteristics, not included to the initial selection criteria.

According to the requirements mentioned in the related legal documents regulating overall evaluation procedure (see: http://www.sciencemon.ru/legal/acts/postanovlenie-pravitelstva-rf-ot-1-noyabrya-2013-979/), the leadership within the reference group is distributed as follows:

- 1st category (leading institutions) = median + 25%
- 3rd category (loosing research functions) = median – 25%
- 2nd category (stable research organisations) = others

General distribution of organisation by reference groups and performance categories is provided below as a possible solution obtained from the available sample and implementation of the above-mentioned criterion. To distinguish between 2nd and 3rd performance categories within the VIII research profile it was suggested to use additional criterion measured as a share of private funding in the intramural R&D expenditure. The role of private financing was discussed in (Coccia, 2004). In other words if an organization demonstrates poor performance level, but not "at public expense" (the share of private funding in total R&D expenditure is equal or above the median level within the relevant field of science) it falls into the 2nd category. Otherwise, it is associated with institutions loosing their research functions and requires further peer-review based expertise.

**Preliminary results**

Achieved sample allowed identification of eight research profiles in six fields of science, which provides adequate representation of reference groups and thereby – the statistical significance of the derived threshold values. In the present structure natural sciences, medical and agricultural sciences as well as technology studies are well represented. Somewhat lower concentration of organizations is seen in in the social sciences and humanities (table 2).

Each research profile is characterized by a unique configuration of resources and results that can not apply the same evaluation criteria and thresholds to the whole population of organizations. In particular, the resulting distribution naturally include confluent profiles of technology developers and service providers (VI) in the field of humanities. Similar misbalances can be seen in other fields.

**Table 1. Model distribution of PRIs by reference groups and performance categories**
Different fields of science demonstrate different functional roles. In particular, in the field of medical sciences most of the organizations appear in the group of technology developers (II and IV profiles, mostly first and second performance categories), agricultural sciences also dominated by the developers as well (specifically in the first category). Social sciences are represented mostly in the VIII profile that is characterized by a high number of ulterior (unexpressed) results mostly due to the high number of libraries and infrastructure elements historically holding status of research organizations. The latter is also true for the technical and technological and natural science areas having a number of pilot plants, experimental stations, etc. The second reason for over-representation of the VIII profile is multiple research orientation by fields of science and inability within the existing framework to divide measurable types of outputs by corresponding fields of science, which is one of the limitations of the proposed approach.

**Discussion and further research**

The proposed approach for reference group identification and assessment could be easily replicated in different samples. However, it has several significant limitations.

The results of the described distribution are largely determined by the boundary values set for the each reference group. Changing these values may lead to significant changes in the whole picture. In the proposed model the thresholds are set automatically, depending on the behavior of all scientific organizations in a sample. Accordingly, any change in the sample can make a difference in the assessment for an individual organization. Such changes may be both productive (e.g. increased performance indicators of some organizations over time can worsen the situation for others) or counterproductive (for instance, through the introduction of unusual objects in the sample for which the proposed indicators would not be relevant).

Secondly, the method does not exclude the possibility for an organization to change its category due to minimal change in one of the performance indicators. In other words, when two organizations in the same field of science demonstrate about the same level of performance and one of them lies on the threshold separating research profiles, then they may be included in different categories that is logically incorrect, taking into account their almost complete identity. Thus, the organizations holding leadership positions in the same pure profiles (I to III) may lose their advantage in relation to one of the mixed profiles (IV to VII), i.e. improving value of an additional indicator, it may lose its leading position by another one.
Third, the current approach does not allow further differentiation of R&D outputs by field of science. Therefore, performance is measured by the overall output of an organisation that may significantly distort the assessment results in case of its multidisciplinary specialization. Further approach might include development of approaches for bibliometric data verification and application of fractional counting for more accurate assessments.

In the end, it should be mentioned that, such an exercise is rather declarative and should be considered only as a preliminary and rough assessment the results of which though may be used for further detailed analysis and peer-review bases evaluation. Further research is aimed at overcoming the described constraints and searching for an optimized solution.

References


Measuring knowledge exchange of internationally mobile scientists -
A bibliometric approach based on similarity

Valeria Aman
German Centre for Higher Education Research and Science Studies (DZHW)
Schützenstraβe 6a, 10117 Berlin
Phone: +49(0)30-2064177-40
aman@dzhw.eu

Abstract
This study introduces a new bibliometric approach to study the effects of international scientific mobility on knowledge exchange. It is based on an analysis of internationally mobile and non-mobile German scientists publishing in journals that are indexed in Scopus. Using bibliometric data such co-authored articles, keywords and references from the Scopus database, I present a technique based on cosine similarity to measure structures of scientific communication. This quantifiable method is capable of revealing patterns of knowledge exchange between internationally mobile scientists and the absorptive capacity of the receiving research groups. Analyses are presented for an overall 9-year publication period (2007 to 2015), split into a pre-mobility phase, a mobility phase and a post-mobility phase, each of which lasts for three years. I compare internationally mobile scientists with non-mobile scientists and discuss the potentials and limitations of the method presented. It is concluded that the bibliometric approach proposed is useful when applied on a large scale to test for significance. International mobility proves to benefit the exchange of ideas between scientists from different countries.

Conference Topic
Methods and Techniques
Studies on the level of individual scientists

Introduction
The internationalization of science systems makes international mobility (IM) relevant to dynamics of knowledge production and knowledge exchange. IM is supposed to facilitate the acquisition and recombination of knowledge that is located at distant places (Laudel, 2003; Gläser, 2006) and is indispensable when the knowledge to be transferred is tacit and thus requires personal contact, observation of colleagues, or informal meetings (Collins, 1974, 2001). It is therefore crucial to understand how IM influences the exchange of scientific knowledge across national borders.

In this study, international mobility of scientists is understood as the physical transition between research organizations that are located in different countries of the world. Cross-border mobility is a means to gain access to international research groups (Ackers, 2008). Previous literature related to geographic mobility and exchange of knowledge focused mainly on the size and direction of migratory flows, neglecting the mechanisms underlying knowledge exchange. In contrast, Ackers suggests that the “focus should shift towards the quality of flows and the nature of knowledge transfer processes” (2005, p.116). Moreover, there is no systematic body of empirical knowledge that allows examining the implications of mobility for the dynamics of the production and diffusion of knowledge (Cañibano et al., 2008). Most previous studies mainly consider patents, investments, and spin-offs as mechanisms of knowledge transfer (Fernández-Zubieta et al., 2015). These formal communication channels are especially common in studies of technology and knowledge transfer. Patent citations were, for instance, used as a proxy for knowledge transfer and frontier innovation by Gibson and McKenzie (2014). Although the circulation of knowledge
at the transnational level is often desired and facilitated by organizational support or job change, the relationship between international mobility and knowledge circulation is relatively under-researched and the field of study suffers from a lack of empirical results. Studying star scientists (authors of highly-cited journal articles) Trippl (2013) also argued that little is known about the nature of knowledge flows of mobile scientists. Most of the existing studies are speculations about the impact of international mobility on knowledge production and transfer, whereas published studies of true knowledge transfer are rare. However, Collins (1974) described a case where an innovative laser could be constructed by scientists who had seen the original set-up and interacted with the constructors, but could not be replicated by scientists who had only read about it and studied diagrams.

This study aims to update our knowledge on the role of international mobility in knowledge exchange processes, by trying to contribute to empirical results that are based on bibliometrics. Bibliometrics can not only be used for institutional evaluations and individual research assessments, but also for measuring intellectual influences of scholarly activities. One of the first studies exploring bibliometric methods related to international mobility was published by Laudel (2003). Additional bibliometric analyses of scientific mobility focused on a defined group of scientists, e.g. doctorate holders (Gaulé & Piacentini, 2013) or the overall population of scientists of a set of countries (Conchi & Michels, 2014; Moed & Halevi, 2014). International scientific mobility can be traced by looking at scientists’ affiliations through the years and their scientific output throughout their career. There are no bibliometric methods yet to indicate the exchange of knowledge. The approach presented in this study aims to provide a first understanding of how internationally mobile scientists engage in knowledge exchange mechanisms. The study examines German scientists and their co-authors publishing between 2007 and 2015 in journals that are covered by Scopus. The scientists under study are not German by nationality but by publishing mainly from German institutions. Internationally mobile German scientists are compared with non-mobile scientists in different scientific fields to account for epistemic characteristics. The cosine similarity of German scientists and their co-authors in publication practice is compared. Salton and McGill (1983) were the first to introduce the cosine between two vectors A and B into information science. The large data set used in this study allows for statistical tests and robust findings. In the rest of the paper, I present a theoretical background that is linked to bibliometrics and explain the database and method.

**Theoretical background and bibliometrics**

Knowledge exchange cuts across a variety of academic fields and application contexts and therefore lacks a common definition; it especially depends on how knowledge itself is defined (Edler et al., 2011). Whereas the term *knowledge transfer* implies that knowledge has been transferred from A to B (and possibly got lost at place A), the term *knowledge exchange* is considered a more appropriate term to describe the exchange of knowledge between (at least) two individuals. In contrast to information that is codified and stored in publications, software or manuals, knowledge is less tangible, has a tacit component and is stored in people’s heads. Assuming that knowledge is embedded in people (Bozeman et al., 2001) the physical interaction and involvement of the knowledge carrier is crucial for any knowledge exchange. Although science invests considerably in codification of knowledge in form of scientific publications, a large part of the knowledge produced remains tacit. Seminal empirical studies on the role of tacit knowledge were published by Collins (1974, 2001) arguing that the transfer of tacit knowledge can only be solved by physical mobility in form of exchange of visits.

Knowledge exchange is tied to knowledge production which is a collective endeavor (Polanyi, 1962) taking place at the level of the individual, the research group and the scientific community. Scientific communities extend over the globe and their members produce new
knowledge claims by interpreting existing knowledge and identifying research gaps that they address in their work (Laudel & Gläser, 2008). Most research efforts result in publications that are available to the scientific community. Peer-reviewed publications assure a minimum level of quality and provide detailed information on the content (title, abstract, and keywords), influential work (references), scientists and organizations, grants, and the geographical location (affiliation). Thus, scientific publications are not only the elementary unit of formal scientific communication, but they also constitute a crucial element in knowledge exchange mechanisms.

The main assumption of this study is that the bibliometric database Scopus covers the core communication channel of some scientific communities. Thus, every active participant contributes to the common body of knowledge by publishing and referencing knowledge claims (Gläser & Laudel, 2001). Journals constitute a major part in the scientific publishing process. Authors choose deliberately appropriate journals to submit their work to and have a specific journal profile that may change over their career.

International mobility is often considered as an integral part of an academic career. Mobility episodes are sequential throughout a career trajectory and can be tracked on the basis of the publication history of a scientist. Finally, research cooperation can describe two different types of co-authorships. Co-authorship can result from the work of a local group of scientists or from a cooperation network that is translocal and relatively stable over time. Both types of co-authorships enable knowledge exchange. The integration of internationally mobile scientists into research groups abroad can manifest itself in co-authorships. Therefore, knowledge exchange is operationalized as co-authored publications which result from formal as well as informal communication and interaction.

Data and methods

Data base
The data were retrieved from Scopus (Elsevier) that is licensed as custom data at the Competence Centre for Bibliometrics and is integrated in an in-house database. A scientist is specified as an author ID in Scopus. The author ID is supposed to combine all publications of an author under a single ID to handle common first and last names (Moed et al., 2013). In order to associate an author with his publication oeuvre, the algorithm uses a multifaceted approach where name spelling variants, affiliations, co-authors, subject areas and the prior publication history are taken into account. The algorithm aims at higher levels of precision than recall. Thus, as soon as a publication cannot be assigned to an existing author ID, a new profile will be created under which the publication appears. The algorithm is supplemented by an author feedback system where an author can indicate whether publications are wrongly attributed to his profile. Only the author ID enables large-scale analysis of international mobility, because it is imprecise to disentangle scientists simply on the basis of their first and last name.

Moed et al. (2013) and Conchi and Michels (2014) published on the reliability of Scopus’ author ID to trace scientists’ mobility and found it acceptable if applied on a large data set so that certain flaws are counterbalanced. Kawashima and Tomizawa (2015) evaluated the accuracy of Scopus’ author ID on the basis of the largest funding database in Japan that provides a unique researcher number. Recall and precision of Scopus author ID for Japanese researchers were reported as 98% and 99%, respectively, showing that the author ID is sufficiently reliable.

To test the precision of the Scopus author ID for German scientists, CVs of all 193 Leibniz-laureates (1999-2016) were coded. The Leibniz Prize is the most important research prize in

---

Germany and approx. ten scientists are awarded each year for their outstanding work. Having compared the laureates’ publication country with the country of residence as derived from CV data it can be concluded that author ID has a high degree of precision for these authors. However, this set is not representative as it includes authors who have a large set of publications. The attribution of a new publication is more precise if it is compared with a large set of previous publications to define it as a similar publication that must belong to the laureate. The ideal case was found for the majority of laureates: a scientist publishing in sources that are covered by Scopus has a single author ID and each author ID relates to a single scientist. Thus, the validation process proved that the author ID algorithm produces good results for German authors.

Internationally mobile scientists are specified as those whose affiliation changes from one country to another. A country relates to the geographic location of the institute at which a scientist conducted and published his work. It is not the country of their nationality or official country of residence.

Erroneous attributions affect not only the accuracy of mobility data but also those on co-authors. Co-authors may be affected by profiles containing papers that in reality belong to other authors. The more common a name, the more papers will be assigned to the profile, which is especially true for Asian scientists (Moed et al., 2013). A co-author’s oeuvre might be split between different profiles due to diverse metadata of publications. As a result, one profile will contain the majority of publications and several smaller profiles only 1-3 papers (Moed & Halevi, 2014). The study takes the potential errors into account by setting restrictions on the data.

Data selection

In a first step, all authors were selected that have at least one paper published with a German affiliation between 2007 and 2015 (n=47,814). Publications are limited to articles and reviews in journals only. German authors are defined as those who have at least half of their publications published with a German affiliation.

The publication period is limited to 2007 to 2015. I distinguish three phases: a pre-mobility phase that ranges from 2007 to 2009, a mobility phase of an undefined duration between 2010 to 2012 and a post-mobility phase that lasts from 2013 to 2015. The mobility episode is characterized as a phase where at least one publication is affiliated to a non-German institution. The duration of the international mobility phase is not defined. Due to Scopus limitations, mobility is measured on a yearly basis (Moed & Halevi, 2014). By international mobility I consequently refer to stays abroad that last from one to three years according to publication data. All of the German scientists under study publish in the phase prior to international mobility as well as after the mobility episode exclusively from German institutions.

Only those German scientists are under study who have published between 6 and 200 papers. By this limitation I try to avoid merged or split identities and also exclude those authors with few publications, because they hardly provide information on how authors move from one country to another. All these restrictions result in 8,371 German scientists.

Co-authors under study are specified as those who have at least two joint publications in at least one of the three phases with German scientists. This guarantees that a co-author and a German author were interacting and not affiliated on the basis of a large collaboration. To exclude errors of homonyms (one author name relating to different persons – common name) and synonyms of co-authors (scientists having more than one author name – split identity), these are limited to those who publish between 3 and 200 publications (2007 to 2015) and are

---

2 A single author may be affected by both types of error.
affiliated at most to 4 countries. Thus, the number of German scientists who co-published with co-authors as described diminishes to 2,134.

As a comparison group, I selected a random sample of 10,000 German scientists who have exclusively published with a German affiliation in the period 2007 to 2015. To measure effects of co-authorship I also distinguish the three phases described above (2007-2009, 2010-2012 and 2013-2015). The same restrictions to co-authors were set as to the group of German scientists who have been abroad. Thus, the number of German scientists in the comparison group diminishes to 2,487.

Methods

The main methodological question is to what extent bibliometrics is capable of indicating knowledge exchange. As stated above, knowledge exchange is operationalized as co-authored publications. Since knowledge exchange manifests itself in the content of the publications of the recipients of knowledge, it is possible to see whether the same sources, keywords and references are used for publications.

To take epistemic practices into account, the authors are grouped by the research field they belong to. This classification scheme is necessary to compare scientists among their scientific peers. Overall, there are 26 research fields distinguished in Scopus’ All Science Journal Classification (AJSC).

Five different publication profiles are considered for each German scientist under study:

(I) The first publication profile contains all publications of the scientist in the post-mobility phase.

(II) The second publication profile contains all publications of the scientist in the pre-mobility phase.

(III) - (V) The other three publication profiles contain all pre-mobility phase publications of the scientists’ co-authors in the pre-mobility phase (III), the mobility phase (IV) and the post-mobility phase (V).

The similarity of German scientists and their co-authors was computed on the basis of the journals. Therefore, the vectors of the numbers of articles per journal in the respective publication profiles are considered. For two vectors A and B, the angle $\alpha$ between them is computed using the following formula:

$$\cos(\alpha) = \frac{A \cdot B}{\|A\| \cdot \|B\|}.$$ 

Cosine similarity is defined as the cosine of the angle of two vectors. The smaller the angle the bigger is the similarity. Figure 1 illustrates this notion of similarity for the publication practice of two authors. The closer the vectors the more similar authors are in their choice of journals. A paired t-test was used to the test the significance of the differences in similarities between publication profile (I) and (II), (I) and (III), (I) and (IV) and (I) and (V).
Results

The following Table 1 provides an overview of the angles between the vectors of number of publications by journal for German authors being internationally mobile (mobile group) and those who were not mobile (comparison group). Only 20 of the 26 AJSC fields are displayed where the number of scientists is at least 80. Note that scientists publish in journals that are assigned to more than one scientific field.

For most disciplines, the similarity of the authors’ post-mobility phase publication profile (I) to the post-mobility co-authors pre-mobility phase publication profile (V) is significantly higher than the similarity of the authors’ post-mobility phase publication profile (I) to their own pre-mobility phase publication profile. This is true both for the internationally mobile German scientists as well as the non-mobile German scientists. Results are inconclusive for the following disciplines: Chemical Engineering (both groups), Energy (both groups), Engineering (both groups) and Materials Science (both groups). Interestingly, the exceptions are Mathematics and Computer Science. In these disciplines German scientists are more similar in the post-mobility phase to themselves in the pre-mobility phase than to any of the co-authors in the pre-mobility phase.

For the disciplines Agricultural and Biological Sciences (both groups), Biochemistry and Molecular Biology (both groups), Immunology and Microbiology (both groups), Medicine (both groups), Neuroscience (both groups), Physics and Astronomy (comparison group), Psychology, Veterinary (comparison group), the similarity of the authors’ post-mobility phase publication profile (I) to the mobility co-authors pre-mobility phase publication profile (IV) is significantly higher than the similarity of the authors’ post-mobility phase publication profile (I) to their own pre-mobility phase publication profile. This is true both for the internationally mobile German scientists as well as the non-mobile German scientists. Results are inconclusive for the following disciplines: Chemical Engineering (both groups), Chemistry (both groups), Dentistry (both groups), Earth and Planetary Sciences (both groups), Environmental Science (both groups), Materials Science (both groups) and Pharmacology (both groups). Again, in Computer Science (both groups), Engineering (both groups) and Mathematics (both groups) German scientists are more similar to their pre-mobility publication profile than to the pre-mobility phase of the co-authors of the mobility phase.
Table 1. Overview of the angles between the vectors of number of publications by journal for German authors being internationally mobile and those who were not mobile. N is the number of scientists in each of the two groups. Column A: Similarity of post-mobility phase (I) to pre-mobility phase (II). B: Similarity of post-mobility phase (I) to co-authors from pre-mobility phase (III). C: Similarity of post-mobility phase (I) to co-authors from mobility phase (IV). D: Similarity of post-mobility phase (I) to co-authors in post-mobility phase (V).

<table>
<thead>
<tr>
<th>Scientific field</th>
<th>N</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>p(A ≠B)</th>
<th>p(A≠C)</th>
<th>p(A≠D)</th>
<th>N</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>p(A ≠B)</th>
<th>p(A≠C)</th>
<th>p(A≠D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agricultural and Biological Sciences</td>
<td>909</td>
<td>73.4</td>
<td>72.6</td>
<td>72.2</td>
<td>69.2</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>930</td>
<td>76.0</td>
<td>75.2</td>
<td>74.7</td>
<td>71.3</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Biochemistry and Molecular Biology</td>
<td>1226</td>
<td>72.3</td>
<td>71.7</td>
<td>71.1</td>
<td>68.5</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>1294</td>
<td>75.4</td>
<td>74.6</td>
<td>74.0</td>
<td>71.0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Chemical Engineering</td>
<td>324</td>
<td>68.1</td>
<td>69.4</td>
<td>68.7</td>
<td>67.2</td>
<td>0.01</td>
<td>0.37</td>
<td>0.23</td>
<td>322</td>
<td>69.1</td>
<td>70.3</td>
<td>70.6</td>
<td>68.1</td>
<td>0.05</td>
<td>0.05</td>
<td>0.21</td>
</tr>
<tr>
<td>Chemistry</td>
<td>719</td>
<td>68.3</td>
<td>68.9</td>
<td>68.5</td>
<td>66.2</td>
<td>0.09</td>
<td>0.77</td>
<td>0.00</td>
<td>734</td>
<td>70.8</td>
<td>71.3</td>
<td>71.6</td>
<td>68.5</td>
<td>0.17</td>
<td>0.10</td>
<td>0.00</td>
</tr>
<tr>
<td>Computer Science</td>
<td>162</td>
<td>66.5</td>
<td>70.4</td>
<td>71.1</td>
<td>68.8</td>
<td>0.00</td>
<td>0.00</td>
<td>0.03</td>
<td>168</td>
<td>69.6</td>
<td>73.8</td>
<td>75.3</td>
<td>71.9</td>
<td>0.00</td>
<td>0.00</td>
<td>0.08</td>
</tr>
<tr>
<td>Dentistry</td>
<td>112</td>
<td>68.9</td>
<td>68.5</td>
<td>68.7</td>
<td>65.4</td>
<td>0.69</td>
<td>0.86</td>
<td>0.00</td>
<td>114</td>
<td>75.4</td>
<td>74.9</td>
<td>75.1</td>
<td>71.2</td>
<td>0.58</td>
<td>0.76</td>
<td>0.00</td>
</tr>
<tr>
<td>Earth and Planetary Sciences</td>
<td>119</td>
<td>66.4</td>
<td>66.2</td>
<td>64.7</td>
<td>61.3</td>
<td>0.89</td>
<td>0.13</td>
<td>0.00</td>
<td>143</td>
<td>69.3</td>
<td>70.1</td>
<td>69.8</td>
<td>67.2</td>
<td>0.36</td>
<td>0.58</td>
<td>0.06</td>
</tr>
<tr>
<td>Energy</td>
<td>165</td>
<td>67.9</td>
<td>68.4</td>
<td>68.0</td>
<td>66.3</td>
<td>0.47</td>
<td>0.89</td>
<td>0.11</td>
<td>145</td>
<td>68.8</td>
<td>70.1</td>
<td>71.8</td>
<td>69.1</td>
<td>0.15</td>
<td>0.00</td>
<td>0.80</td>
</tr>
<tr>
<td>Engineering</td>
<td>451</td>
<td>65.2</td>
<td>67.9</td>
<td>67.5</td>
<td>65.7</td>
<td>0.00</td>
<td>0.00</td>
<td>0.46</td>
<td>486</td>
<td>68.8</td>
<td>70.9</td>
<td>71.8</td>
<td>70.1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.07</td>
</tr>
<tr>
<td>Environmental Science</td>
<td>300</td>
<td>69.3</td>
<td>70.1</td>
<td>70.2</td>
<td>66.7</td>
<td>0.13</td>
<td>0.18</td>
<td>0.00</td>
<td>320</td>
<td>71.5</td>
<td>72.2</td>
<td>72.1</td>
<td>68.1</td>
<td>0.24</td>
<td>0.42</td>
<td>0.00</td>
</tr>
<tr>
<td>Immunology and Microbiology</td>
<td>303</td>
<td>72.6</td>
<td>71.5</td>
<td>71.3</td>
<td>67.8</td>
<td>0.04</td>
<td>0.04</td>
<td>0.00</td>
<td>296</td>
<td>76.0</td>
<td>75.2</td>
<td>73.6</td>
<td>71.0</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Materials Science</td>
<td>477</td>
<td>65.4</td>
<td>66.9</td>
<td>65.9</td>
<td>64.2</td>
<td>0.00</td>
<td>0.41</td>
<td>0.09</td>
<td>548</td>
<td>68.9</td>
<td>69.9</td>
<td>69.9</td>
<td>67.8</td>
<td>0.03</td>
<td>0.06</td>
<td>0.08</td>
</tr>
<tr>
<td>Mathematics</td>
<td>97</td>
<td>65.0</td>
<td>68.6</td>
<td>70.3</td>
<td>67.9</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>80</td>
<td>69.4</td>
<td>72.2</td>
<td>73.7</td>
<td>71.4</td>
<td>0.03</td>
<td>0.00</td>
<td>0.24</td>
</tr>
<tr>
<td>Medicine</td>
<td>1678</td>
<td>71.7</td>
<td>71.0</td>
<td>70.4</td>
<td>68.0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1919</td>
<td>75.3</td>
<td>74.1</td>
<td>73.7</td>
<td>70.5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Neuroscience</td>
<td>580</td>
<td>73.4</td>
<td>72.5</td>
<td>71.9</td>
<td>69.4</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>634</td>
<td>77.2</td>
<td>76.0</td>
<td>75.1</td>
<td>71.7</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Nursing</td>
<td>184</td>
<td>71.6</td>
<td>71.9</td>
<td>70.8</td>
<td>68.6</td>
<td>0.68</td>
<td>0.31</td>
<td>0.00</td>
<td>175</td>
<td>74.7</td>
<td>73.8</td>
<td>72.3</td>
<td>70.2</td>
<td>0.16</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Pharmacology</td>
<td>258</td>
<td>70.3</td>
<td>69.4</td>
<td>69.3</td>
<td>67.4</td>
<td>0.10</td>
<td>0.11</td>
<td>0.00</td>
<td>213</td>
<td>75.8</td>
<td>75.8</td>
<td>75.6</td>
<td>71.1</td>
<td>0.92</td>
<td>0.75</td>
<td>0.00</td>
</tr>
<tr>
<td>Physics and Astronomy</td>
<td>684</td>
<td>63.7</td>
<td>64.9</td>
<td>63.8</td>
<td>61.8</td>
<td>0.01</td>
<td>0.92</td>
<td>0.00</td>
<td>690</td>
<td>66.9</td>
<td>68.1</td>
<td>68.2</td>
<td>65.7</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>Psychology</td>
<td>200</td>
<td>72.0</td>
<td>72.0</td>
<td>71.7</td>
<td>69.7</td>
<td>0.88</td>
<td>0.64</td>
<td>0.00</td>
<td>190</td>
<td>77.2</td>
<td>76.2</td>
<td>74.8</td>
<td>72.3</td>
<td>0.15</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Veterinary</td>
<td>153</td>
<td>72.3</td>
<td>71.5</td>
<td>71.9</td>
<td>69.7</td>
<td>0.20</td>
<td>0.62</td>
<td>0.01</td>
<td>177</td>
<td>76.2</td>
<td>75.2</td>
<td>74.0</td>
<td>70.4</td>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
For the disciplines Agricultural and Biological Sciences (both groups), Biochemistry and Molecular Biology (both groups), Immunology and Microbiology (mobile groups), Medicine (both groups) and Neuroscience (both groups) the similarity of the authors’ post-mobility phase publication profile (I) to the pre-mobility co-authors pre-mobility phase publication profile (III) is significantly higher than the similarity of the authors’ post-mobility phase publication profile (I) to their own pre-mobility phase publication profile. 

In Chemical Engineering (mobile group), Computer Science (both groups), Engineering (both groups), Materials Science (both groups), Mathematics (both groups) and Physics and Astronomy (both groups), German scientists are significantly more similar to their pre-mobility phase publication profile than to the pre-mobility publication profile of the co-authors in the pre-mobility phase. Results are inconclusive for the following disciplines: Chemistry (both group), Dentistry (both groups), Earth and Planetary Sciences (both groups), Energy (both groups), Environmental Science (both groups), Nursing (both groups), Pharmacology (both groups), Psychology (both groups) and Veterinary (both groups).

Figure 2 illustrates the cumulative distribution of angles between pre- and post mobility publication profiles of internationally mobile scientists and non-mobile scientists. 

The distribution of angles is computed over all scientists and is not distinguished by disciplines. It is evident that internationally mobile scientists are more similar to themselves than non-mobile scientists. Non-mobile scientists show a high share (ca. 30%) of post-mobility publication profiles that are (journal-)orthogonal to their 2007-2009 publication profiles, i.e. in the 2013-2015 phase they publish in other journals than in the 2007-2009 phase. In comparison, 20% of internationally mobile scientists publish in the post-mobility phase in completely different journals than in the pre-mobility phase. The degree of the angle between the pre- and post-mobility publication profiles of the internationally mobile scientists is 70.05, whereas the angle of the non-mobile scientists is 73.29.

Assuming that publications sharing the same keywords result from interaction or a common language, knowledge exchange was also operationalized as the shared keywords and references of German scientists and their co-authors Thus, the vectors of the number of
articles by keyword and the number of articles by reference in the respective publication profiles were considered. The closer the vectors the more similar authors are in their choice of keywords and references. The tables are not listed due to space limitations but the results are overall similar to the results presented on the basis of journals. However, the angles are higher because the number of keywords and references to be used is higher than the limited number of journals covered in Scopus.

**Discussion**

The study aimed at measuring the impact of international mobility on publication practices, where knowledge exchange was operationalized as co-authorships. The study revealed that in some disciplines the similarity of German authors to their co-authors increases with international mobility during the period of investigation, whereas in other disciplines authors remain similar to themselves.

In general, the degree of similarity is higher for the mobile scientists in comparison to those without a mobility phase between 2010 and 2012. Evidently, scientific mobility is more than meetings between individuals. When scientists meet in person, they can screen each other’s skills and adjust to each other making the knowledge exchange more efficient. Especially, disciplines in which the research work is often conducted outside the national borders of a country and scientific findings result from interaction triggered by international mobility and collaboration, an increased similarity of the publication practices were measured. In contrast, in Mathematics and Computer Science where the knowledge is mostly codified, no positive impact of co-authors during the mobility phase is measurable.

The methodology presented does not aim at assessing the performance of individual scientists but rather the effects of international mobility at higher aggregation levels. The analysis of large datasets and the choice of a comparison group tend to cancel out randomly distributed errors. However, systematic errors are not accounted for, because certain limitations are inherent to bibliometric studies of international mobility. Bibliometric research allows tracking mobility only to the extent that scientists publish and that the affiliation is stated on the publication and is linked to them. Another factor influencing results is that a stay abroad does not necessarily result in a publication, and thus cannot be measured as a stay abroad. Moreover, the publication delay has to be kept in mind, thus a mobility episode that becomes evident between 2010-2012 on the basis of publication from this phase, may result from earlier stays abroad.

Another challenge to the bibliometric approach is that it is based on the use of unique author profiles that rely on the Scopus algorithm (Moed et al. 2013). Additional distortions may derive from scientists who move to another country but continue to publish with colleagues from their old institution and from double affiliations, thus when scientists are simultaneously appointed at institutions that are located in different countries. A limited coverage of the database can also cause inaccuracies in detecting mobility episodes.

I tried to minimize the limitations by restricting the period to recent years (2007-2015) because the Scopus author ID works best for recent publications. Moreover, on the basis of German authors whose residence countries (as provided in their CVs) were compared against the country codes in the affiliations of their publications, I made sure that the precision of the author ID is sufficiently high. Additional restrictions were set in regard to the publication counts and the co-authors.

**Conclusion**

The mere inflow and outflow rates of mobile scientists are insufficient to studies of the impact of international mobility on knowledge aspects. The need to study the effects of international mobility on knowledge exchange constitutes new challenges for empirical research. The aim
was to provide insight into the uses and limits of the new approach. The author ID and the affiliation information in Scopus made it possible to distinguish German scientists that were internationally mobile from those who were not. Using a bibliometric approach that is based on affiliation data in scientific publications it is possible to draw some conclusions regarding the similarity of authors and their co-authors abroad as well as in their country of residence. Thus, it can be concluded that bibliometrics can provide insight into research processes and indicate knowledge exchange and that the pilot study provides significant outcomes in terms of similarity of mobile authors and their co-authors abroad.

However, the methodology presented is only able to show patterns of similarity on the publication practice and covers the specific content of the knowledge, thus the semantic relationship of research terms (i.e. keywords), only to some degree. The topical orientation accompanying knowledge exchange can be analyzed on the basis of abstracts and language processing techniques such as co-word analyses. Thus, it remains important to develop tools for better understanding the nature of knowledge flows of mobile scientists and the value added that international mobility brings to the individual scientist as well as to research. In the current study, every journal is treated as completely distinct from every other journal. Future work should consider the similarity of journals and incorporate this similarity into the calculation of the similarity of publication profiles.

Finally, it is important to be keep in mind that co-authorships can stand for different relationships and therefore need to be distinguished in future work. Co-authorships can either occur because scientists are diametric in their research focus and complement each other’s work or because they are similar in their topical orientation and work on the same issue.

Acknowledgments
The study was funded by the Bundesministerium für Bildung und Forschung (BMBF) under the grant number 01PQ16002. The data builds on the bibliometric database provided by the Competence Centre for Bibliometrics (grant number: 01PQ17001).

References


Journals as Communication Channels between National Sub-Communities

Jochen Gläser\(^1\) and Valeria Aman\(^2\)

\(^1\)Jochen.Glaser@ztg.tu-berlin.de
TU Berlin (Germany)

\(^2\)aman@dzhw.eu
German Centre for Higher Education Research and Science Studies (DZHW) Hanover (Germany)

Abstract

Most uses and investigations of scholarly journals rest on implicit understandings of journals as communication channels which are rarely made explicit, let alone discussed. In this paper, we argue that by theoretically embedding and operationalizing ‘internationality’, bibliometrics can contribute to the study of social structures of scientific communities. We consider journals as communication channels in the production process of scientific communities and define their internationality as the links they create between community members from different countries. This definition can be operationalised with four sets of variables describing the international composition of the set of authors, the selection processes and their impact on the international composition of the set of authors, the international composition of the journal’s audience, and knowledge flows between authors and audiences. We analyse six journals from different regions of the world in the field of International Relations and find that the supposedly ‘leading international journals’ are in fact closer to being domestic US-journals, while European and Asian journals show a much higher diversity of authors, audiences and editorial boards. We confirm that a differentiated analysis of international diversity contributes to theory-driven investigations of scientific communities. Our approach appears to be transferable to questions about the thematic diversity of journals.

Conference Topic
Indicators
Methods and Techniques
Science of Science

Introduction

Scholarly journals are one of the core empirical objects of bibliometrics. Journal-to-journal citations are used for the mapping of scientific fields (Leydesdorff 1987) and the construction of science overlay maps (Leydesdorff and Rafols 2009). Journal-level citation indicators such as journal impact factors are critically assessed, and suggestions for further developments are made (Moed et al. 1996; Glänzel and Moed 2002). In addition, a variety of properties of journals including gatekeeping functions (Braun and Bujdosó 1983; Ni et al. 2013) and internationality (Zitt and Bassecoulard 1999; Aman 2016; Andrei et al. 2016) are studied. These uses and investigations of scholarly journals rest on implicit understandings of journals as communication channels which are rarely made explicit, let alone discussed. In this paper, we argue that by paying attention to theories of scientific production and the role of journals in this production process, bibliometric studies of journals can make major contributions to science studies. By theoretically embedding and operationalising the idea that ‘internationality’ is a theoretically important property of journals, we illustrate possible contributions by bibliometrics to the study of social structures of scientific communities. We thereby take on the task formulated by Leydesdorff (1989: 339) nearly thirty years ago:

*Only by further specification of what empirical studies contribute to specific theoretical questions can we systematically further the relations between qualitative theory and scientometric methods in S&T-studies.*
Theory

Journals as communication channels

A common weakness of both the Mertonian and the constructivist sociology of science is their implicit separation of a local construction and subsequent certification and circulation of contributions to scientific knowledge. This separation is explicit in Zuckerman’s and Merton’s (1971) consideration of the peer review as a process of certifying knowledge. Despite their explicit rejection of Merton’s structural-functionalist approach, constructivist science studies implicitly kept this distinction by strongly focusing on micro-social processes of knowledge construction, a focus which included a relative neglect of wider knowledge structures and the role of published scientific knowledge in the laboratory (Lynch 1985: 153).

If we start from the idea that scientific knowledge is produced collectively by scientific communities, scientific communication must be reconceptualised as a part of the production process rather than a process of circulating products. In this production process, individual publications are offers of intermediate contributions to a common product, which might or might not be accepted by further use in subsequent knowledge production processes. Changes in communication practices re-structure a community’s knowledge production process.

From this perspective, a scholarly journal can be defined as a formally organised communication channel that contributes to maintaining the collective production of scientific knowledge by periodically making publicly accessible selected scientific contributions for use by others. The selection process is under control of members of scientific communities who form an editorial board and provide peer reviews. In most cases, the communication channel is not internal to one scientific community but makes accessible contributions from several communities for the use by several (not necessarily the same) communities, as is reflected by Bradford’s law (Brookes 1969).

The communicative function of journals consist in linking specific contributions to specific audiences. This function varies in several dimensions. The most important dimension for science studies is content, i.e. the knowledge flows through journals. These knowledge flows are produced by gatekeeper decisions on scope, quality standards, publication language, and other properties of the journal. The perception of these decisions and their implementation in the day-to-day selection of contributions shape self-selection processes among potential authors and decisions to become part of the journal’s audience. By providing selected content from a self-selected (and subsequently externally selected) set of authors to a self-selected set of readers, journals contribute to the emergence of content-based sub-communities, i.e. community members who address the same topic.

Another dimension that is relevant to the study of scientific communities, which we will operationalise and investigate in more detail in this paper, is the internationality of a journal. The ‘internationality of journals’ has enjoyed considerable attention, partly because it is considered an aspect of a journal’s quality (Buela-Casal et al. 2006) and partly because it reflects the communicative function of linking national sub-communities to their international communities (Aydinli and Mathews 2000; Andrei et al. 2016). Unfortunately, the concept is usually taken for granted and not theoretically defined. The attempt by Buela-Casal et al. (2006) to derive an understanding from encyclopaedia entries illustrates the problem, namely the absence of a theory in which the concept is embedded.

1 Actor-network-theory and its account of fact-making as a collective process (Latour 1987: 29) may look like an exception but still leaves out the process that turns isolated facts into scientific knowledge.

2 From other scientific perspectives, scholarly journals could be defined differently, i.e. as products of a commercial publisher (Volkmann et al. 2014).
If we consider journals as communication channels in the production process of scientific communities, the internationality of a journal can be defined as \textit{the link it creates between community members from different countries}. This link is rather complex and can be described by four sets of variables, namely
- variables describing the international composition of the set of authors;
- variables describing the selection processes and their impact on the international composition of the set of authors;
- variables describing the international composition of the journal’s audience; and
- variables describing knowledge flows between authors and audiences. Similar to content-based communities, journals can create national or regional sub-communities if self-selections and selections are based on nationality or region.

The internationality of journals in the field of International Relations

The field of International Relations has discussed its own international relations for a long time. Beginning with Hoffman’s (1995 [1977]) “An American Social Science: International Relations”, members of the IR community have examined the degree to which their field includes contributions from countries beyond the US or the North American and Western European core. Contributions to this discussion are often based on the analysis of journal publications, which play an important role in IR. For example, Aydinli and Mathews (2000) studied the dynamics of core-periphery relationships in IR by analysing the authorship of 20 leading IR journals in the 1990s. They constructed four categories of countries (core of the core, periphery of the core, core of the periphery, and periphery of the periphery) and found proportions of publications from these groups to be constant during the 1990s. Turton (2016) analyses contributions to 12 IR journals in an attempt to answer the question whether scholars from the US dominate the field of IR. Her investigation goes beyond bibliometrics by categorising article content. She did not find much evidence supporting the thesis of US dominance. However, her conceptualisation of dominance as intentional or at least conscious process (ibid, 11-12) seems unduly narrow because it does not pay enough attention to predominance as an emergent effect of individual actions.

This discussion and its empirical referents can be framed as the sociological question of the internal structuration of scientific communities. The sociology of science has analysed the differentiation of scientific communities in center and periphery (e.g. Crane 1972; Chu 1992a; b; Hwang 2008) as well as the role of language (Brittain 1984) and of nationally specific research objects (Reguant and Casadella 1994; Rey-Rocha and Martin-Sempere 1999) for the emergence of national sub-communities. However, empirical studies have been separated from theory, and measures have rarely been operationalisations of theoretical concepts. The IR field appears to be ideally suited for a study of sub-structures because it utilizes journal-based communications but at the same time features nationally or regionally specific research objects, a variety of publication languages, and theoretical pluralism. In the context of a larger project on the internal structuration of and knowledge flows in the IR field, the question arose how the internationality of journals and possible internal boundaries created by communicative functions of journals could be explored. In order to contribute to this investigation, we operationalised the concept of internationality and tested our indicators on six journals in the field.

\footnote{A short description of the project can be found at https://www.researchgate.net/project/Overlapping-Scientific-Communities-Internal-Structuration-and-Knowledge-Diffusion-in-International-Relations.}
Data and Methods

Data

We test our approach by applying it to six journals (Table 1). From the discussion of core-periphery relationships in IR follows that the internationality can be expected to vary between US-based journals and journals in other regions. The population from which journals can be selected is limited due to biases of the Web of Science. We therefore included one journal indexed in the SciELO database, whose records can also be downloaded from WoS in the same format. Five of the journals publish in English. We used all document types. Most of the documents in Revista Contexto Internacional (176) were published in Portuguese, 20 documents were published in English.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese Journal of International Politics</td>
<td>CJIP</td>
<td>China</td>
<td>Asia</td>
<td>137</td>
</tr>
<tr>
<td>Revista Contexto Internacional</td>
<td>RCI</td>
<td>Brazil</td>
<td>South America</td>
<td>196</td>
</tr>
<tr>
<td>European Journal of International Relations</td>
<td>EJIR</td>
<td>UK</td>
<td>Europe</td>
<td>320</td>
</tr>
<tr>
<td>International Organization</td>
<td>IO</td>
<td>USA</td>
<td>North America</td>
<td>283</td>
</tr>
<tr>
<td>International Relations of the Asia-Pacific</td>
<td>IRAP</td>
<td>Japan</td>
<td>Asia</td>
<td>214</td>
</tr>
<tr>
<td>International Studies Quarterly</td>
<td>ISQ</td>
<td>USA</td>
<td>North America</td>
<td>563</td>
</tr>
</tbody>
</table>

The data originate from the bibliometric database of the German Centre for Bibliometrics, which builds upon licensed custom data files of Thomson Reuters’ Web of Science (WoS). The journal Revista Contexto Internacional is not part of WoS’ Core Collection and is thus not included in our licensed version of WoS data. We therefore downloaded the full records manually from the SciELO Citation Index that is implemented in WoS. The data were edited (e.g. coding the country codes of authors and citers) and integrated in the local database. The publication period ranges from 2006 to 2015.

Since IR scholars, and thus potential authors and readers, are unevenly distributed across countries, distributions observed for the six journals must be assessed against reference values. We construct the baseline from all journals in WoS that are assigned to the Subject Category ‘International Relations’ and are covered in WoS for at least five years within the period 2006-2015 (91 journals). The reference set contains authors from a total of 151 countries and citers from 170 countries.

Data about editorial board membership of the six journals were downloaded from the internet (online versions of journals and journal websites).

Methods

We operationalise the international composition of a journal’s author set as the number of addresses from each country. If authors provide addresses from more than one country, all countries are counted equally. Countries are counted fractionally, i.e. a value of 0.25 is added to each country if an article is linked to addresses from four countries.

The international composition of a journal’s audience is operationalised as the number of addresses from each country for citing articles. This operationalisation is problematic for two...
reasons. First, citers constitute only as small proportion of a journal’s audience. Second, the nationality of citers can be determined only for the source items, i.e. for those citing articles that are indexed in the WoS. Since IR as a social science relies on several communication channels beside WoS-indexed journals and is ill represented due to the English language bias of the WoS, the citing source items on which the analysis of audience internationality is based constitute only a fraction of citing publications. Taking into account the citation dynamics of the social sciences, citation counts were calculated for a five-year period of the actual year of publication plus the four subsequent years. Thus, for papers published in 2006, citations from the publication period 2006–2011 were considered. This 5-year-citation window allows calculating citation rates of papers published up to 2011. Self-citations were included because they contain important information, namely authors’ belonging to the audience of the journal in which they publish.

Selection processes cannot be analysed bibliometrically. However, some information on selection processes is indirectly provided by looking at ‘regulatory body’ internationality, which we operationalize as international diversity of editorial boards.

The analyses of author sets, audiences and ‘regulatory bodies’ measures several aspects of the international diversity of the three groups of actors. A system’s diversity is commonly understood as the heterogeneity produced by the disparity of its elements (Stirling 2007; Gläser et al. 2015). It can be analysed by categorising the elements of a system and measuring three properties, namely
- variety (the number of categories, here: countries),
- disparity (the dissimilarity of categories, here: countries), and
- evenness (the distribution of elements across categories, here: the number of authors/editorial board members/readers from each country).

Since there is no meaningful measure for disparity of countries, we only use variety and evenness, which are combined in the Shannon Index

\[ I = -\sum_{i=1}^{n} p_i \ln(p_i) \]

where \( I \) is the Shannon Index of the distribution of elements across categories, \( n \) is the number of categories and \( p_i \) is the share of elements in category \( i \). For any number of categories the Shannon Index \( I \) achieves its maximum when all \( p_i \) are all equal (all categories have the same share of elements). The Shannon Index \( I \) drops to 0 when all elements belong to one category.

A disadvantage of synthetic diversity measures such as the Shannon Index is that differences in diversity cannot be easily ascribed to one of the properties. This is why it is useful to look at variety and evenness separately.

In order to measure the internationality of authorship, the Shannon Index was computed for the period 2006 to 2015 on a yearly basis as well as for the whole publication period.

The internationality of knowledge flows is assessed by two variables. First, we measure journal self-citation rates in order to assess the degree to which a communication channel is closed (i.e. author set and audience coincide) at the journal level. Second, we measure country-self-citation rates, i.e. the share of references by authors of a country that refer to publication of their same country. This measure goes beyond the international diversity of author sets and audiences by asking to what extent knowledge flows run across countries (are indeed international) or are just national knowledge flows that co-exist in journals.

Results

Authorship internationality

Table 2 provides an overview of the variety of authorship internationality for the six journals, i.e. the number of countries that had authors in the six journals between 2006 and 2015.
Table 2. Variety of authors’ countries and comparison to baselines, 2006-2015.

<table>
<thead>
<tr>
<th>Journal</th>
<th>No. of countries</th>
<th>% of 91 journal baseline (151)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese Journal of International. Politics</td>
<td>18</td>
<td>12.77</td>
</tr>
<tr>
<td>Revista Contexto Internacional</td>
<td>23</td>
<td>16.31</td>
</tr>
<tr>
<td>European Journal of International Relations</td>
<td>31</td>
<td>21.99</td>
</tr>
<tr>
<td>International Organization</td>
<td>18</td>
<td>12.77</td>
</tr>
<tr>
<td>International Relations of the Asia-Pacific</td>
<td>24</td>
<td>17.02</td>
</tr>
<tr>
<td>International Studies Quarterly</td>
<td>27</td>
<td>19.15</td>
</tr>
</tbody>
</table>

The evenness of authorship internationality is described by Table 3, which lists the distribution of authorships across countries for the ten most frequently occurring countries. The dominant authoring countries are the publishing countries shown in Table 1.

Table 3. Distribution of authorship across the 10 most occurring countries for each journal and the baseline of 91 journals, 2006-2015.

<table>
<thead>
<tr>
<th>CJIP</th>
<th>RCI</th>
<th>EJIR</th>
<th>IR</th>
<th>IRAP</th>
<th>ISQ</th>
<th>91 Journal baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHN</td>
<td>36.4</td>
<td>BRA</td>
<td>70.7</td>
<td>GBR</td>
<td>28.6</td>
<td>USA 82.2 JPN 29.4 USA 72.3 USA 33.5</td>
</tr>
<tr>
<td>USA</td>
<td>27.7</td>
<td>USA</td>
<td>5.9</td>
<td>USA</td>
<td>27.7</td>
<td>GBR 3.2 USA 22.8 GBR 10.6 GBR 20.3</td>
</tr>
<tr>
<td>GBR</td>
<td>7.9</td>
<td>GBR</td>
<td>4.9</td>
<td>GER</td>
<td>6.9</td>
<td>CAN 2.6 GBR 13.5 CAN 2.8 GER 5.0</td>
</tr>
<tr>
<td>AUS</td>
<td>6.2</td>
<td>ARG</td>
<td>2.9</td>
<td>AUS</td>
<td>6.6</td>
<td>GER 2.3 AUS 6.6 AUS 2.2 CAN 4.9</td>
</tr>
<tr>
<td>GER</td>
<td>3.3</td>
<td>CAN</td>
<td>2.0</td>
<td>CAN</td>
<td>4.2</td>
<td>CHE 2.0 KOR 5.5 NOR 1.9 AUS 4.6</td>
</tr>
<tr>
<td>SGP</td>
<td>2.9</td>
<td>PRT</td>
<td>2.0</td>
<td>SWE</td>
<td>3.7</td>
<td>AUS 1.4 SGP 4.7 GER 1.8 CHN 2.2</td>
</tr>
<tr>
<td>CAN</td>
<td>2.5</td>
<td>FRA</td>
<td>1.5</td>
<td>NLD</td>
<td>2.4</td>
<td>NOR 1.2 CAN 2.4 DNK 0.9 NOR 2.1</td>
</tr>
<tr>
<td>JPN</td>
<td>2.5</td>
<td>ZAF</td>
<td>1.5</td>
<td>NOR</td>
<td>2.4</td>
<td>SWE 1.0 CHN 2.1 ISR 0.9 KOR 2.1</td>
</tr>
<tr>
<td>TWN</td>
<td>2.1</td>
<td>GER</td>
<td>1.0</td>
<td>DK</td>
<td>2.3</td>
<td>KOR 0.7 MYS 1.9 SWE 0.9 NLD 1.7</td>
</tr>
<tr>
<td>BEL</td>
<td>1.7</td>
<td>ESP</td>
<td>1.0</td>
<td>FIN</td>
<td>2.1</td>
<td>CHN 0.6 TWN 1.9 NLD 0.9 TUR 1.5</td>
</tr>
<tr>
<td>Sum</td>
<td>93.0</td>
<td>93.2</td>
<td>86.9</td>
<td>97.1</td>
<td>90.8</td>
<td>95.2</td>
</tr>
</tbody>
</table>

The proportions of countries with the highest shares of authors in a journal’s total output range between 29% and 82%, and considerably deviate from the proportions of the 91 journal baseline. These striking differences already suggest influences of publication language, topics and national sub-communities.

Table 4 lists the Shannon Index for the six journals for the whole period, which must be compared to the baseline of I= 1.88 for the 91 journals.

Table 4. Authorship diversity (Shannon Index), 2006-2015.

<table>
<thead>
<tr>
<th>Journal</th>
<th>Shannon Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese J. of Int. Politics</td>
<td>1.97</td>
</tr>
<tr>
<td>Revista Contexto Internacional</td>
<td>1.28</td>
</tr>
<tr>
<td>European J. of Int. Relations</td>
<td>2.31</td>
</tr>
<tr>
<td>Int. Organization</td>
<td>0.91</td>
</tr>
<tr>
<td>Int. Relations of the Asia-Pacific</td>
<td>2.22</td>
</tr>
<tr>
<td>Int. Studies Quarterly</td>
<td>1.24</td>
</tr>
</tbody>
</table>
Figure 1 shows that authorship diversity is slightly increasing over time for all journals and the baselines with the exception of the *Chinese Journal of International Politics*, whose diversity markedly increases and moves it from the group of lower-diversity journals to the group of higher-diversity journals.

![Internality of authorship graph](image)

**Figure 1. Shannon diversity of authorship for the six journals and the 91 journals baseline, 2006-2015.**

**Citer internationality**

The variety of citer internationality is provided in Table 5. Table 5 also lists the share of journal self-citations, which indicates the degree to which a journal provides a ‘closed’ communication circuit. Given the low share of self-citations for all journals, none of them can be considered ‘closed’. Finally, the table lists the proportion of source items in the journals’ references on which these and the following data are based. These must be compared to an overall share of 25.05% source items in all 91 IR journals.

**Table 5. Variety of citers’ countries, comparison to baseline, journal self-citations, and share of source items, 2006-2011.**

<table>
<thead>
<tr>
<th>Journal</th>
<th>Number of countries</th>
<th>% of 91 journal baseline (170)</th>
<th>Proportion of source items in references</th>
</tr>
</thead>
<tbody>
<tr>
<td>CJIP</td>
<td>24</td>
<td>14.12</td>
<td>18.25</td>
</tr>
<tr>
<td>RCI&lt;sup&gt;5&lt;/sup&gt;</td>
<td>2</td>
<td>1.18</td>
<td></td>
</tr>
<tr>
<td>EJIR</td>
<td>48</td>
<td>28.24</td>
<td>29.94</td>
</tr>
<tr>
<td>IO</td>
<td>47</td>
<td>27.65</td>
<td>45.86</td>
</tr>
<tr>
<td>IRAP</td>
<td>32</td>
<td>18.82</td>
<td>23.70</td>
</tr>
<tr>
<td>ISQ</td>
<td>56</td>
<td>32.94</td>
<td>42.84</td>
</tr>
</tbody>
</table>

<sup>5</sup> Due to the different data set, proportions of journal self-citations and of source items in *Contexto Internacional* could not be analysed.
The evenness of citer distribution across countries is provided in Table 6. Again, a baseline needs to be considered because the citations received depend on the country distribution of readers, which is best estimated by author distributions because authors are potential citers. Table 6 shows a higher evenness of citer distributions than author distributions for all journals except Revista Contexto Internacional. The extreme values for Revista Contexto Internacional are easily explained by the predominant publication language. Between 2006 and 2011, all articles were published in Portuguese.

The predominance of citing countries coincides with the predominance of authoring countries for four journals, while Chinese Journal of International Politics and International Relations of the Asia-Pacific were mostly cited by US scholars, who are not the most frequent authors of these journals. Although the audiences appear to be dominated by citers from the US, the attention of US authors is clearly focused on the US journals, which have more than 50% citers from the US compared to the baseline of 32%.

Table 6. Distribution of citer nationalities across the 10 most occurring countries for each journal and the baseline of 91 journals, 2006-2015.

<table>
<thead>
<tr>
<th>Journal</th>
<th>CJIIP</th>
<th>RCI</th>
<th>EJIR</th>
<th>IO</th>
<th>IRAP</th>
<th>ISQ</th>
<th>91 journals baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>21.1</td>
<td>BRA</td>
<td>96.0</td>
<td>GBR</td>
<td>24.2</td>
<td>USA</td>
<td>59.9 USA</td>
</tr>
<tr>
<td>CHN</td>
<td>18.9</td>
<td>PRT</td>
<td>4.0</td>
<td>USA</td>
<td>21.2</td>
<td>GBR</td>
<td>8.9 JPN</td>
</tr>
<tr>
<td>GBR</td>
<td>15.9</td>
<td></td>
<td></td>
<td></td>
<td>8.4</td>
<td></td>
<td>7.0 GBR</td>
</tr>
<tr>
<td>AUS</td>
<td>8.4</td>
<td>CAN</td>
<td>7.3</td>
<td>CAN</td>
<td>4.1</td>
<td></td>
<td>11.6 CAN</td>
</tr>
<tr>
<td>JPN</td>
<td>6.6</td>
<td></td>
<td>AUS</td>
<td>6.3</td>
<td>CHE</td>
<td>3.1</td>
<td>KOR 5.4 AUS</td>
</tr>
<tr>
<td>DNK</td>
<td>5.1</td>
<td></td>
<td>DNK</td>
<td>4.2</td>
<td>AUS</td>
<td>2.0</td>
<td>GER 5.1 NOR</td>
</tr>
<tr>
<td>GER</td>
<td>4.5</td>
<td></td>
<td>SWE</td>
<td>3.5</td>
<td>NOR</td>
<td>1.7</td>
<td>CAN 4.5 SWE</td>
</tr>
<tr>
<td>KOR</td>
<td>3.6</td>
<td></td>
<td>NLD</td>
<td>2.8</td>
<td>SWE</td>
<td>1.4</td>
<td>CHN 3.9 CHE</td>
</tr>
<tr>
<td>SWE</td>
<td>3.0</td>
<td></td>
<td>CHE</td>
<td>2.6</td>
<td>DNK</td>
<td>1.3</td>
<td>TWN 2.9 ISR</td>
</tr>
<tr>
<td>CAN</td>
<td>2.8</td>
<td></td>
<td>ISR</td>
<td>2.2</td>
<td>ITA</td>
<td>1.2</td>
<td>SGP 2.8 DNK</td>
</tr>
</tbody>
</table>

Sum 89.9 100.0 82.5 90.6 85.6 89.0 76.0

Table 7 lists the Shannon Index for the citer audiences of the six journals for the whole period from 2006 to 2011, which must be compared to the baseline of I= 2.24 for the 91 journals.

Table 7. Audience diversity (Shannon Index), 2006-2011 (5-year citation window).

<table>
<thead>
<tr>
<th>Journal</th>
<th>Shannon Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese J. of Int. Politics</td>
<td>2.41</td>
</tr>
<tr>
<td>Revista Contexto Internacional</td>
<td>0.17</td>
</tr>
<tr>
<td>European J. of Int. Relations</td>
<td>2.61</td>
</tr>
<tr>
<td>Int. Organization</td>
<td>1.77</td>
</tr>
<tr>
<td>Int. Relations of the Asia-Pacific</td>
<td>2.58</td>
</tr>
<tr>
<td>Int. Studies Quarterly</td>
<td>1.98</td>
</tr>
</tbody>
</table>

Figure 2 illustrates how the internationality of citing authors developed over the years. Overall, the internationality increased over the years. Again, the Chinese Journal of International Politics shows a marked increase of diversity and moves from the group of lower-diversity journals to the group of higher-diversity journals. The low values for Revista
Contexto Internacional are likely due to the publication language - almost all citers are from Brazil.

Figure 2. Overview of the Internationality of citing countries for the six journals of interest and the average journal internationality of 91 IR-journals in the publication period 2006-2011 and a citation window of 5 years.

'Regulatory body' internationality
Differences and changes in author internationality would be explained by self-selections and selection processes, which is why it is important to look at the composition of the group of scholars who make decisions on the acceptance of submissions. We focus on the international composition of editorial boards (editorial committees in the cases of The European Journal of International Relations and Revista Contexto Internacional). Table 8 provides data on the diversity of these boards/committees.


<table>
<thead>
<tr>
<th>Journal</th>
<th>Size of editorial board</th>
<th>No. of countries</th>
<th>Dominant country (share in %)</th>
<th>Shannon Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese J. of Int. Politics</td>
<td>20</td>
<td>10</td>
<td>China (25.00)</td>
<td>2.11</td>
</tr>
<tr>
<td>Revista Contexto Internacional (2017)</td>
<td>4</td>
<td>3</td>
<td>Brazil (75.00)</td>
<td>0.95</td>
</tr>
<tr>
<td>European J. of Int. Relations (2017)</td>
<td>20</td>
<td>9</td>
<td>UK (25.00)</td>
<td>1.96</td>
</tr>
<tr>
<td>Int. Organization</td>
<td>39</td>
<td>5</td>
<td>USA (38.30)</td>
<td>2.12</td>
</tr>
<tr>
<td>Int. Relations of the Asia-Pacific</td>
<td>47</td>
<td>16</td>
<td>USA (58.57)</td>
<td>1.71</td>
</tr>
</tbody>
</table>

Knowledge flow internationality
We analyse two variables describing knowledge flows. Table 9 provides journal self-citation rates for the five journals for which they could be calculated. Journal self-citation rates indicate the extent to which journals constitute closed communication circuits. None of the
journals is closed in this sense, a large majority of citations links the journals to other communication channels. This is not surprising given the large proportion of non-source items in IR journals. WoS journals are only one of several important channels of communication in IR.


<table>
<thead>
<tr>
<th>Journal</th>
<th>Proportion of journal self-citations</th>
<th>Proportion of source items in references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese J. of Int. Politics</td>
<td>2.83</td>
<td>18.25</td>
</tr>
<tr>
<td>European J. of Int. Relations</td>
<td>7.78</td>
<td>29.94</td>
</tr>
<tr>
<td>Int. Organization</td>
<td>2.00</td>
<td>45.86</td>
</tr>
<tr>
<td>Int. Relations of the Asia-Pacific</td>
<td>0.73</td>
<td>23.70</td>
</tr>
<tr>
<td>Int. Studies Quarterly</td>
<td>1.23</td>
<td>42.84</td>
</tr>
</tbody>
</table>

Table 10 summarizes the country self-citation rates of the five WoS journals calculated for papers published between 2006 and 2015. In order to avoid distortions by low numbers of references, we selected the ten countries with the most references from each journal. The country self-citation rate expresses the share of references to publications from the author’s country. For example, papers published by authors from the US in CJIP contain 73.6% references to publications from other US authors. Country self-citation rates must be compared to the country-self citation rate among all 91 journals, which is 37.77%.

Table 10. Country self-citation rates for the 10 most frequently occurring countries in citations of the five WoS journals.

<table>
<thead>
<tr>
<th>CJIP</th>
<th>EJIR</th>
<th>IO</th>
<th>IRAP</th>
<th>ISQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>73.6</td>
<td>USA</td>
<td>83.3</td>
<td>70.0</td>
</tr>
<tr>
<td>GBR</td>
<td>23.2</td>
<td>GBR</td>
<td>26.3</td>
<td>17.5</td>
</tr>
<tr>
<td>SWE</td>
<td>13.6</td>
<td>ISR</td>
<td>14.7</td>
<td>10.0</td>
</tr>
<tr>
<td>AUS</td>
<td>9.5</td>
<td>GER</td>
<td>10.0</td>
<td>9.2</td>
</tr>
<tr>
<td>SGP</td>
<td>7.5</td>
<td>AUS</td>
<td>6.9</td>
<td>8.5</td>
</tr>
<tr>
<td>CHN</td>
<td>5.6</td>
<td>SWE</td>
<td>6.1</td>
<td>5.9</td>
</tr>
<tr>
<td>CAN</td>
<td>4.5</td>
<td>NLD</td>
<td>5.9</td>
<td>5.0</td>
</tr>
<tr>
<td>JPN</td>
<td>3.0</td>
<td>DNK</td>
<td>5.6</td>
<td>2.0</td>
</tr>
<tr>
<td>KOR</td>
<td>2.4</td>
<td>CAN</td>
<td>4.0</td>
<td>1.7</td>
</tr>
<tr>
<td>GER</td>
<td>0.9</td>
<td>AUT</td>
<td>1.0</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Discussion

Despite the severe limitation by the low proportions of source items in IR journals, our analysis exemplifies the advantage of a multidimensional analysis of internationality. First, it suggests that some allegedly international or regional journals are better considered domestic journals because scholars from the same country are predominant among authors and audiences (low variety, evenness, and Shannon diversity). This applies to International Organization and International Studies Quarterly, which are both considered “leading international journals” in the field but effectively channel communications between US scholars. The higher audience diversities indicate that the international community observes these communications, but country self-citation rates show that scholars from outside the US
are not really part of them. Revista Contexto Internacional also appears to be a domestic journal. Its diversity appears to be constrained due to the publication language. These three journals are also those whose editorial boards are dominated by one country and show the lowest Shannon diversities in our sample, which points to the impact of editorial policies. These editorial policies must not necessarily target the selection of authors from particular countries but can include preferences concerning empirical objects, theories or methods. Thus, author internationality might be an unintended side-effect of editorial policies on other matters.

The Chinese Journal of International Politics shows an interesting dynamics. Both its authorship diversity and its audience diversity have markedly increased between 2006 and 2015, and it moved from the group of ‘domestic’ to the group of ‘international’ journals in this period. The diversity of the international board in 2015 suggests that this might be due to a policy to increase diversity.

High country self-citation rates point to internal communication of national sub-communities through the analysed journals. This internal communication behaviour is particularly strong in the case of the US national sub-community. In some cases national sub-communities either don’t exist or communicate through other channels.

Conclusions
Due to space limitations, we could only provide a ‘proof of concept’ of our approach. Nevertheless, our analysis enables three conclusions.

First, bibliometric analyses can provide a nuanced picture of journal internationality that supports the study of internal structuration of international scientific communities. As is usually the case, the identification of interesting patterns by bibliometrics raises questions that cannot be answered by bibliometrics alone. However, raising these questions and contributing to the triangulation of qualitative methods and data that must be used to answer them is an irreplaceable contribution of bibliometrics to science studies.

Second, our analysis of IR suggests that IR might still be an ‘American discipline’ in the sense of the US national sub-community being predominant (being the largest sub-community) and introverted (one is tempted to say ‘parochial’). It might well be that the field of IR looks completely different to US and non-US scholars, a question that can be further explored by both bibliometric and other methods.

Finally, the approach applied here can be extended to properties of scholarly communication beyond internationality. For example, keywords or content analyses of articles can be used to study the communicative function of journals with regard to topics. In this case, questions of topic diversity, thematic diversity of audiences, and knowledge flows within and between topics could be studied in a similar fashion as internationality.

Acknowledgments
Part of this work was financially supported by the DFG (grant RI 798/11-1).

References


Authors’ Publication Strategies and Citation Distributions in Journals

Andrey Lovakov1  Vladimir Pislyakov2

1lovakov@hse.ru
Center for Institutional Studies, National Research University Higher School of Economics, Moscow (Russia)

2pislyakov@hse.ru
Library, National Research University Higher School of Economics, Moscow (Russia)

Abstract
Psychology is a discipline standing at the crossroads of hard and social sciences. Some of psychology journals are attributed to SCIE in Web of Science database while others to SSCI (and some to both). So it is especially interesting to study bibliometric characteristics of psychology journals. We study not the citedness itself (IF etc.) but the citation distribution across papers within psychology publications. This is, so to say, “indicators of the second order” which measure the digression of the citations received by individual papers from the journal’s average. This also influences the publication strategies of the authors. Some journals guarantee to the author receiving of the mean number of citations while others have much more “All or Nothing” grade when any individual paper may have many cites or not have them at all. We also define four different types of psychology journals and explore their characteristics separately.

Conference Topic
Journals; Databases and electronic publications; Citation and co-citation analysis

Introduction
Citation distributions are generally characterized as “highly skewed”, this means a non-Gaussian pattern and high concentration of citations on some items while others are low-cited or non-cited at all, see papers by Whitehand (1985), Seglen (1992) and many others. Still, the uneven distribution of citations has its own levels of inequality and may be measured by common indicators such as Gini index, which will be used in the present paper.

The Gini index is a measure of inequality of a distribution. In terms of citation distribution it is a comparison of the journal where all citations are received by one paper with the case when each journal’s paper gets the same number of citations. The Gini index ranges from 0 to 1 where zero corresponds to perfect equality (each paper receives the same number of citations) and 1 corresponds to perfect inequality (only one paper has citations while others are not cited at all).

We are interested in how different journals gain their average citedness per paper. Some may get citedness from close to even distribution of citations, when each article has a high probability to receive an average number of cites. Other publications get their averages from a limited number of highly cited papers, while all others are not cited. The same research question was raised by Nuti et al. (2015) for cardiovascular journals. We choose psychology journals to investigate the issue. Psychology is chosen as a discipline standing at the crossroads of hard and social sciences, which makes it especially interesting for the analysis.

Gini index is used to measure the inequalities of citation distributions by Chi (2016) where the unevenness of book citations is compared with journals and the strong influence of uncited items on Gini coefficient is observed. Wren (2016) plainly states that “The higher Gini coefficients for bioinformatics journals suggest that development of novel bioinformatics resources may be somewhat ‘hit or miss’. That is, some approaches become widely adopted and produce a disproportionate number of extremely highly cited papers, while most are not widely adopted <...>.”
Gini coefficients for citations to a particular journal were also studied by, for example, Chatterjee, Ghosh & Chakrabarti (2016) and by Stegmann & Grohmann (2001). The latter paper analyzes journals on dermatology and explores the dependence of Gini index from journal impact factor. We address this question in the present paper too. Gini indices for leading multidisciplinary journals (Science, Nature, PNAS etc.) and for PhysRev series may be found in (Ghosh, Chattopadhyay & Chakrabarti, 2014).

Gini measure of inequality is also used in the other applied informetrics studies. For example Pislyakov (2008) applies this indicator to usage statistics of electronic journals. Leydesdorff & Rafols (2011) use it, along with other inequality indicators, to measure interdisciplinarity of a journal. They also agree that "The Gini coefficient <...> has the advantage of having been widely used in bibliometrics".

The new thing we want to introduce is the connection between the citation patterns in the journals and the choice the author makes when he/she chooses the appropriate journal to submit his/her manuscript. The choice is whether an author prefers to securely receive the average level of citations which is common to the journal (the case of publications with uniformly distributed citations across papers) or he/she wishes to risk and submit the paper to a journal where just several papers get extremely high level of citedness while others are being under-cited.

This choice marks different publication strategies of a scientist (whether chosen consciously or not). “All is safe” vs. “All or Nothing”.

Data and Methods

Data on 40 psychological journals from Social Sciences Citation Index (SSCI, Web of Science, Clarivate Analytics, ex-Thomson Reuters) were used (Table 1). We limit our analysis only to psychological journals, so there is no problem of cross-disciplinary comparison. However, we distinguish four types of journals in psychology, which, as we will further observe, demonstrate different patterns of citation inequality.

We have identified four types of psychological journals: multidisciplinary, sub-disciplinary, journals dedicated to a single particular topic, and methodological publications. In Web of Science there is another classification schema, with ‘Psychology: Experimental’, ‘Psychology: Developmental’ and others (8 different categories), but for the present research we need the classification ranging across all the sub-disciplines of psychology but marking the specific type of a journal.

Multidisciplinary group of journals publishes research on all spectrum of the science of psychology. For example, Psychological Science publishes papers on topics ranging from cognitive, social, developmental, and health psychology to behavioral neuroscience and biopsychology. Sub-disciplinary group includes journals which cover only one area of psychology. For example, Journal of Personality and Social Psychology publishes papers in areas of personality and social psychology while Cognitive Psychology publishes papers about attention, memory, language processing, perception, problem solving, and thinking. The group of journals dedicated to a single particular topic is focused on some special research area. For example, Depression & Anxiety focuses on studies related to different aspects of mood and anxiety disorders and related phenomena, and Personal Relationships publishes studies focused on attributes of individual partners in personal relationships, interactive relationship processes, and relationships in social contexts. Finally, methodological group includes journals which are...
devoted to research design, methodology, measurement, quantitative and qualitative data analysis. For example, *Behavior Research Methods* publishes papers on methods, techniques, and instrumentation of research in experimental psychology, and *Psychological Methods* on methods for collecting, analyzing, understanding, and interpreting psychological data. The sample includes different journals in terms of impact factor and publisher.

Five-year journal impact-factors (2015) for each journal were extracted from Journal Citation Reports (JCR, Web of Science, Clarivate Analytics, ex-Thomson Reuters). Five Gini indexes were calculated for each journal. Every Gini index was based on citations of papers published in each single year, from 2010 to 2014 (only “articles” and “reviews” document types were taken into calculation). These five indexes were then aggregated using the arithmetic mean. The five-year 2015 impact is used as the same papers are assessed by it: citations of articles/reviews from 2010 to 2014 are included in the formulae of impact-factor.

For the data preparation, analysis and visualization we used R, a programming environment for statistical computing (R Core Team, 2016). The Gini indexes were calculated by the *reldist* package (Handcock, 2016). Additional packages were used: *dplyr* (Wickham & Francois, 2016), *ggplot2* (Wickham, 2009), *readxl* (Wickham, 2016).

### Results and Discussion

Table 1 shows Gini index for 40 psychological journals chosen for our analysis. The Gini varies from 0.41 to 0.64 proving that there are different citation patterns, from close to “All or Nothing” as in *Psichothema* or *Behavior Research Methods* to “All is safe” as in *Psychology of Men and Masculinity*.

If we plot the dependence of Gini index from IF for all 40 analyzed journals, no clear regularity is observed (Figure 1). It seems that journal’s citedness has no evident connection with the distribution of citations across its papers.

But when psychological publications are subdivided into classes by types of journals, the relationship of inequality of citation distribution to impact factor becomes more obvious. We may see it in Figure 2, where linear regression lines with 95% confidence intervals for predictions from a linear model are shown.

What is interesting, and what is the main point of our paper, this has a clear connection to a publication strategy of an author. For example, if you publish in sub-disciplinary journals, the more cited journal you choose, the more “safety” in receiving average citations you get (Pearson’s $r = -0.87, p < 0.001$). But when an author submits his/her paper to a multidisciplinary journal, he/she may choose several ones with the same IF but with significantly different risk of receiving/non-receiving the average number of citations (so to say “All or Nothing” grade).

There is different level of inequality of citation distribution between groups of journals. Gini indexes of all journals from the methodological group are higher than 0.5, whereas Gini indexes of only a couple of journals from sub-disciplinary and special topic groups exceeds 0.5. It means that in general the inequality of citation distribution is higher in journals, which publish papers on methodology, methods, and tools of psychological research. These results are in accord with Wren (2016) that bioinformatics journals, which publish top papers, have higher Gini indexes. Publishing on methodology and research methods in psychology seems also to be 'hit or miss' in terms of citedness. The probable explanation may be that some methodological papers attract attention of many researchers who will use and cite them because these methods are well-
received by psychologists and are becoming widespread. While other papers are not so interesting and popular being rather complex or addressing too specific problems. It is in the methodological part of the science the inequality is therefore becomes more pronounced.

Table 1. Average Gini index and 5-year impact factor for the psychological journals.

<table>
<thead>
<tr>
<th>Journal</th>
<th>Average Gini index</th>
<th>5-year impact factor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multidisciplinary</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychological Science</td>
<td>0.46</td>
<td>6.29</td>
</tr>
<tr>
<td>British Journal of Psychology</td>
<td>0.52</td>
<td>3.26</td>
</tr>
<tr>
<td>Frontiers in Psychology</td>
<td>0.51</td>
<td>2.88</td>
</tr>
<tr>
<td>Acta Psychologica</td>
<td>0.51</td>
<td>2.41</td>
</tr>
<tr>
<td>Journal of Psychology</td>
<td>0.48</td>
<td>1.76</td>
</tr>
<tr>
<td>Scandinavian Journal of Psychology</td>
<td>0.52</td>
<td>1.55</td>
</tr>
<tr>
<td>International Journal of Psychology</td>
<td>0.56</td>
<td>1.42</td>
</tr>
<tr>
<td>Australian Journal of Psychology</td>
<td>0.45</td>
<td>1.26</td>
</tr>
<tr>
<td>Psicothema</td>
<td>0.64</td>
<td>1.23</td>
</tr>
<tr>
<td>Spanish Journal of Psychology</td>
<td>0.58</td>
<td>0.84</td>
</tr>
<tr>
<td><strong>Sub-disciplinary</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Journal of Personality and Social Psychology</td>
<td>0.42</td>
<td>7.44</td>
</tr>
<tr>
<td>Child Development</td>
<td>0.47</td>
<td>5.81</td>
</tr>
<tr>
<td>Cognitive Psychology</td>
<td>0.46</td>
<td>5.49</td>
</tr>
<tr>
<td>Journal of Personality</td>
<td>0.47</td>
<td>4.18</td>
</tr>
<tr>
<td>Memory &amp; Cognition</td>
<td>0.48</td>
<td>2.59</td>
</tr>
<tr>
<td>Personality and Individual Differences</td>
<td>0.48</td>
<td>2.42</td>
</tr>
<tr>
<td>Journal of Cross-Cultural Psychology</td>
<td>0.50</td>
<td>2.24</td>
</tr>
<tr>
<td>Journal of Economic Psychology</td>
<td>0.50</td>
<td>2.06</td>
</tr>
<tr>
<td>Small Group Research</td>
<td>0.48</td>
<td>1.32</td>
</tr>
<tr>
<td>Journal of Applied Social Psychology</td>
<td>0.53</td>
<td>1.31</td>
</tr>
<tr>
<td><strong>Methodological</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychological Methods</td>
<td>0.63</td>
<td>9.46</td>
</tr>
<tr>
<td>Organizational Research Methods</td>
<td>0.55</td>
<td>6.51</td>
</tr>
<tr>
<td>Multivariate Behavioral Research</td>
<td>0.62</td>
<td>4.78</td>
</tr>
<tr>
<td>Behavior Research Methods</td>
<td>0.64</td>
<td>3.98</td>
</tr>
<tr>
<td>British Journal of Mathematical and Statistical Psychology</td>
<td>0.61</td>
<td>2.74</td>
</tr>
<tr>
<td>Journal of Mathematical Psychology</td>
<td>0.59</td>
<td>2.61</td>
</tr>
<tr>
<td>Psychometrika</td>
<td>0.54</td>
<td>2.58</td>
</tr>
<tr>
<td>Methodology, European Journal of Research Methods for the Behavioral and Social Sciences</td>
<td>0.56</td>
<td>2.17</td>
</tr>
<tr>
<td>Educational and Psychological Measurement</td>
<td>0.54</td>
<td>1.85</td>
</tr>
<tr>
<td>Journal of Educational Measurement</td>
<td>0.53</td>
<td>1.34</td>
</tr>
<tr>
<td><strong>Special topic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression and Anxiety</td>
<td>0.46</td>
<td>5.52</td>
</tr>
<tr>
<td>Emotion Review</td>
<td>0.49</td>
<td>4.12</td>
</tr>
<tr>
<td>International Journal of Eating Disorders</td>
<td>0.49</td>
<td>3.69</td>
</tr>
<tr>
<td>Autism</td>
<td>0.42</td>
<td>3.48</td>
</tr>
<tr>
<td>Psychology of Addictive Behaviors</td>
<td>0.43</td>
<td>3.42</td>
</tr>
<tr>
<td>Intelligence</td>
<td>0.48</td>
<td>3.29</td>
</tr>
<tr>
<td>Psychology of Men &amp; Masculinity</td>
<td>0.41</td>
<td>3.12</td>
</tr>
<tr>
<td>Journal of Gambling Studies</td>
<td>0.43</td>
<td>2.98</td>
</tr>
<tr>
<td>Personal Relationships</td>
<td>0.47</td>
<td>1.57</td>
</tr>
<tr>
<td>Perception</td>
<td>0.56</td>
<td>1.13</td>
</tr>
</tbody>
</table>
Figure 1. Relationship between average Gini index and 5-year impact factor (the whole sample of journals).

Figure 2. Relationship between average Gini index and 5-year impact factor by types of journals (linear regressions are shown, grey color displays 95% confidence interval for predictions from a linear model).
For instance, the most cited paper from methodological journals in our sample (860 total citations at the moment of data extraction) is an introduction to propensity score methods for observational studies (Austin, 2011) which are relevant not only for psychological research but also for all fields with observational (or nonrandomized) studies. And the third most cited paper (540 citations in total) demonstrates how to use Amazon’s Mechanical Turk web site for conducting behavioral research (Mason & Suri, 2012) which is extremely useful tool for many researchers in psychology.

We may also remind the well-known ‘Acta Crystallographica A case’ when only one methodological paper (Sheldrick, 2008) has changed the impact factor of the whole journal which has reached the top-3 overall (i.e. within all Web of Science) IF ranking. It was more than 20 times IF-rise for this journal. So, the methodological inequalities are common to the science and we may observe this in case of psychological publications too.

**Conclusion**

It is well known in bibliometrics that journal citedness (for example impact factors) should be compared only within the same scientific discipline. Moreover, there are differences between journal types (those which publish more review papers are generally cited more, for example). The present paper shows that the same approach should be used when we study not only average citedness but also patterns of digression of received by individual papers’ citations from this average journal citedness. Psychology journals of different types have different Gini indices and different relationship between Gini and 5-year impact factor.

This clearly marks differences in publication strategies. Of course, when the author chooses the publication venue, at the first place stands the subject matter of the journal and its aptness to his/her research interests. Next, generally goes the international visibility/popularity of the journal which is often proxied by the citedness of the periodical.

But as we may see from our research, another characteristic of the journal also matters. It is so to say the grade of “All or Nothing”. How strong the scientist may be confident in receiving average number of journal’s citations to his/her individual paper. Some journals have more equal citation distribution across their papers than the others with the same IF.

Though often this choice is made not consciously by the authors, they may feel it from empiric impressions from different journals (if they have a scholarly career long enough). Anyhow, the authors should know and keep in mind these “differences of the second order” which may influence their publication strategies.

At last, we should underline that though the average of Gini for five years was taken for the present paper, the further research is needed to examine whether the regularities found are robust over time. Moreover, the effect most probably will vary across scientific disciplines which is also a fascinating topic of future analysis.

**Acknowledgments**

The paper was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE), and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'.
References


Scientometric Research of Knowledge Communication on Social Media
-A Case Study of Biomedical Science on Baidu Baike

Cheng Ni\(^1\) and Dong Ke\(^2\)

\(^1\)ncheng@mail.hzau.edu.cn
Huazhong Agricultural University, Wuhan(China)

\(^2\)dk8047@163.com
Wuhan University, Wuhan(China)

Abstract
Social media stand as perfect examples of Web 2.0 applications. Baidu Baike, a wiki-like online encyclopedia, is analyzed as a typical collaborative project contributed by experts and users. How do the contributors communicate during the edition of entries? In order to seek answers to the question, scientific entries are chosen for citation analysis to reveal relationships among entries, which could indicate communication of entries’ contributors. The relationship between online entries and traditional literatures is discussed on the basis of reference analysis, which indicates communication between entries’ contributors and literatures’ authors. Cooperation analysis among experts or users is also conducted. Analysis on coupling entries helps to indicate the common interests of contributors. The crossing and integration of disciplines was confirmed in this paper. It is found that people’s online communication was affected by traditional knowledge carriers. The majority of experts cooperated with someone in the common affiliation while cross-unit cooperation also existed. There are clusters of common interests of experts, which indicate that the specialist only masters his own field. Moreover, users, who are typical general public, have wide range of interests and full of enthusiasm for entry contribution. Their participation is of great importance to development of science and technology.

Keywords
Citation and co-citation analysis; Co-occurrence analysis; Knowledge Communication; Baidu Baike

Conference Topic
Citation and co-citation analysis; Co-occurrence analysis; Science communication

Introduction
Knowledge management has received growing attention from both researchers and practitioners. With the rapid development of Web 2.0 applications, knowledge communication and sharing has moved beyond face-to-face exchanges to social media contexts such as blogs, Bulletin Board System (BBS), Wikipedia, Facebook, Twitter, and so forth. (Kling & McKim, 2000; Kling, McKim & King, 2003; Hoisl, Aigner & Miksch, 2007; Hsu, et al. 2006; Klamma, Cao & Spaniol, 2007) The feasibility of using social media in scholarly communication has been researched. Both of Koh(2007) and Wagner(2014) pointed out that social media could be used for knowledge sharing and communication. Eijkman(2010) thought that Wikipedia could be used in scholarly communication. Other researchers (Park, 2009; Liu, 2010; Dantonio, 2012) gave affirmative opinions on social media in study, research and cooperation. Moreover, influence of social media on scholarly communication was studied. Collins(2010), Kirkup(2010), and Gruzd(2012) agreed that social media can be used to build connections in order to stimulate new research ideas. Letierce(2010), Tiryakioglu(2011), and Gruzd(2012) found that scholars benefited from social media for academic exchange or scholar communication. See Yin
Lim(2014) indicate that the performance of undergraduate informatics students was more excellent after using social media for academic purpose. Those online encyclopedias, such as Wiki, were considered as valuable tools for knowledge communication and group collaboration (Wagner, 2004; Sauer, 2005). Moskaliuk (2009), Cho (2010), and Begoña (2011) investigated the use and influence of Wiki during knowledge sharing and communication. It is reasonable to consider online encyclopedias as an effective way for scholarly communication though qualitative research methods were used in the past research work mentioned above. However, Wiki is so free that every user can change the content arbitrarily. In order to investigate scholarly knowledge communication on social media, an online wiki-like encyclopedia, Baidu Baike, is chosen. The reason is as follows. Baidu Baike is a free, open, and typical online encyclopedia platform on the Web operated by Baidu Online Network Technology Co., Ltd., whose official version was released on April 21, 2008. More than 13 million entries had been included by April, 2016. And more than 5.8 million people who are on Baidu Baike involved in the edition of entries. Baidu Baike is considered as the world's largest Chinese Encyclopedia. The scholarly entries on Baidu Baike are user-generated and locked after review or modification by experts to keep them scientific and authoritative, which is so different from normal Wikipedia in which only users contribute to the entries. The scholarly entries are reliable and scientific. The scholarly communication on Baidu Baike is so potential that the hyperlinks, historic versions, and references behind will be analyzed. In this paper, the answer to these questions will be discussed. (1) What is the composition of related entries on Baidu Baike? (2) Are online entries related to traditional literature? (3) How did the contributors of entries cooperate together? (4) Is it possible for us to reveal the interest of contributors?

Methods and data collection

Methods

Scientometrics is becoming increasingly important, common, and sophisticated (Bornmann and Daniel, 2008). Scientometrics provides new metrics of scholarly impact on the social Web and useful tools to analyze developments of scientific fields (Rip & Courtial, 1984; Courtial, 1994; Hassan & Loebbecke, 2010; Priem & Hemminger, 2010). Citation analysis and co-words analysis is used to identify research group (Reyes-Gonzalez et al., 2016; Shiau & Dwivedi, 2013). Knowledge flow is investigated with citation analysis and co-word analysis (Tsay, 2015; Lee, 2017; Ravikumar et al., 2015).

Scientometric methods are used to analyze the relationship among the entries, contribution and cooperation from users and experts. Citation analysis is used to reveal communication among citing and cited authors or entry contributors taking hyperlinks and references as pointcuts. Co-word analysis is used to investigate cooperation among entry contributors including users and experts. Besides, it is also used to find out relationship among entries in the field of Biomedical Science. Bibliographic coupling is used to reveal the relationship among entries based on their references and contributors including users and experts. Knowledge aggregation is the act of perspective extraction of information, from a group of information sources based on divergent human perception, reasoning, and learning (Abiodun Robert, 2009).

The methods of knowledge aggregation are used to analyze linked data for knowledge discovery (Jun et al., 2015), survey of websites of academic libraries regarding the adoption of Web 2.0 technologies (Mahmood & Richardson, 2011), and online privacy and reputation management (Portmann & Pedrycz, 2014).
Data Collection

The study focuses on online scholarly entries in the field of Biomedical Science. The related data of entries is collected from official website of "Science" section of BaiduBaike (http://baike.baidu.com/science). And the data is processed through several technologies such as data mining and data cleansing.

There are 251 academic entries in the field of Biomedical Science on BaiduBaike. However, there are some repeated entries including “biomacromolecule mechanics”, “tissue engineering”, and “dental Biomechanics”. In addition, the entry “corneal contact lens” and “contact lenses” are same because they are synonymous exactly. We keep “contact lenses” instead of “corneal contact lens” for the sake of simplicity. Therefore, dereplication is used on the repeated and synonymous entries. There are 247 entries finally. The references, hyperlinks, and historic versions of each entry were collected manually. And information on experts who contributed to the entries was also recorded.

Composition of Related Entries on Baidu Baike

Is every entry isolated? Is there any relationship among entries? Although some entries “stay alone” without any hyperlink, it does not mean they are isolated exactly. The number of hyperlinks in the entry depends on edition of contributors.

It is hyperlinks in entries that indicate that editors of one entry have found, referred and cited other entries during the edition process, which is a typical communication activity. The hyperlinks of each entry on Baidu Baike are other entries. We tried to collect the hyperlinks data of 247 entries mentioned above in order to reveal the relationship among citing and cited entries, which indicates communication among entries contributors.

There are only 40 out of 247 entries which cited other entries. There are 6 entries self-cited including “contact lenses”, “regenerative medicine”, “biological solid mechanics”, “Bioregenerative Life Support System (BLSS)”, “implant denture”, and “Prostate cancer stem cells (PCSCs)”. The entry “Bioregenerative Life Support System (BLSS)” cited itself for 3 times. And there is no citation among these 40 entries. However, they cited other 618 entries. The citation network of entries is shown as Figure 1.

The entry “magnetic resonance” cited 116 entries. It cited entry “magnetic moment” for 18 times, “nuclear magnetic resonance” and “ferromagnet” for 11 times respectively, and “ferromagnetic resonance” for 10 times. The entry “cell cycle” cited 81 entries while “contact lenses” cited 62 ones, “nanomaterials” cited 59, and “color Doppler flow imaging” cited 54.
Each of “Carbon nanotubes”, “tissue engineering”, “Oral CT”, and “catheter” cited more than 20 entries. There are 8 entries which cited only one entry respectively, including “bite force”, “boneresorption”, “Biophotonics”, “respiratory quotient”, “biological solid mechanics”, “nano-biomaterials”, “stasiology”, and “respiratory mechanics”.

It is found that entry contributors put in some hyperlinks citing other related entries in order to explain the entry they were working on. Thus they searched, read, chose and cited some related information from other entry contributors. The process indicates communication among entry contributors beyond limit on time and space.

Analysis of citation among entries indicates the composition of entries online. The hyperlinks cited by the entries in the field of Biomedical Science come from other disciplines such as Chemistry, Physics, Oncology, and so forth. To some extent, it indicates cross and integration of disciplines. There are no borders within Science.

The Relationship between Online Entries and Traditional Literatures

In addition, there are references in some entries to provide readers with more related further reading materials and clues to the literature to help them access the original literature. Analysis on the references of entries is of great help to reveal relationship between online entries and traditional literatures, such as journal articles, books, conferences and so forth.

The 247 entries cited 1252 references in total, including 1026 journal articles, 133 books (monographs and teaching materials), 45 conference literatures, 18 web pages, 15 dissertations, 2 patents, and so on. There are 5 references which are internal handouts or news without clear sources.

There are 11 books that were cited by the entries more than 3 times. For example, Sport Biomechanics written by Hu(2013) was cited for 10 times. Theory and technology of Bioregenerative Life Support System by Liu(2009) was cited for 8 times. Besides, the number of books which were cited for more than 3 times is 9. There are 50 literatures which were cited twice, including 28 journal articles, 17 books, 4 dissertations, and one conference literature. It is seemed that academic journals and monographs were preferred by entry contributors because of their mature theoretical system and novel knowledge which will make the entries more scientific and reliable.

There are some entries which have the mutual references. Entry bibliographic coupling based on references is used to reveal potential relationship among entries and communication among their contributors (shown as Figure 2).
It is found that some entries are linked so closely that some small network or clusters appear. For example, some entries are about life support, including the space life support system, technology, ecosystem, and so forth. Some entries related to biomechanics are clustered, such as ankle biomechanics, hip biomechanics, shoulder joint kinematics, and so on.

The online entries are new form of knowledge but some of their roots come from literatures. It seemed that content-rich entries’ contributors had absorbed academic and scientific knowledge from journals, conferences, books, and even patents. Another way to think about it is the cited literatures’ authors transferred knowledge to these contributors. Their communication was across the time and space.

**Cooperation of Contributors**

It is found that there is no less than one contributor for each entry. Some entries only have one contributor who creates the entries. Some are results of some normal users. The others are contributed by users and experts. The co-occurrence of experts or users is considered the cooperation of them. It is obvious that cooperation is common during the contribution of entries. The cooperation may be among users or experts. It is also possible that users and experts mutually contributed to the entries. According to the historic versions of entries on Baidu Baike, it is quite easy to find out the time and the users who contributed to them. However, some entries were modified by experts without the record of details including the time, the way, and the content. It is difficult to investigate how the users and experts cooperated during the contribution though the cooperation really happened. What we can do is to analyze the cooperation among experts and cooperation among users.

**Cooperation among experts**

There are 206 entries whose contribution with the involvement of 89 experts in all. Each of them involves 1 to 3 experts. There are 96 entries which are results of cooperation of experts. Both of GuHanqing and Wan Rongxin from Tianjin Medical University contributed to 35 entries, followed by NiuWenxin from Tongji University and Fan Yubo from Beijing University of Aeronautics and Astronautics. Besides, there are 6 experts who participate in more than 10 entries respectively. There are 39 experts who only involve in one entry. The cooperation network can be drawn based on occurrence analysis of experts (see Fig.3). The affiliation of experts was analyzed. And it is found that overwhelming majority of experts preferred to cooperating with their workmate or someone who work in the same affiliation. But there are exceptions. For example, cooperation of Tongji University and the 307th Hospital of Chinese People’s Liberation Army, Peking University and Tianjin Medical University, Suzhou University and Hong Kong Polytechnic University can also be found, which break through the limitation of time and space. It is Internet and communication technology that made it true.
Cooperation among users

The historic versions of each entry, which showed the created time, modified time and participated users, were analyzed in order to find out who had contributed to the entries. The number of historic versions of the entry “contact lenses” is up to 296. There are 236 users who contributed to the entry, in which there is one user who participate the edition of this entry for 18 times.

There are 115 entries which are results of cooperation of users while 132 entries were contributed by the user named “csbme” alone. There are 1379 users who had participated in contribution of these 115 entries. The cooperation network of users (see Figure 4) is the result of analysis on historic versions of each entry. There are some active users who contributed to entries for many times and cooperated with others a lot. For example, the user named “fodaoshuangxiu(佛道双休)” participated in the edition for 441 times and cooperated with the user named “matalvshenglingshe(马他吕圣灵社)” for 315 times, with the user named “yashangyougedan(牙上有个蛋)” for 294 times. The user “baike ROBOT(百科 ROBOT)” cooperated with the user “w_ou” for 97 times, and with the user “csbme” for 68 times. It is found that the potential connection of these users is built based on the occurrence of the users of each entry. The users would participate in some edition because of their common interest in some entries. Although the users did never met one another, they mutually contributed to the entries through the online platform.
The cooperation among experts or users reveals the potential communication online, in which experts or users may not know each other, especially users’ true identity is hidden in the virtual environment. However, communication and cooperation did happen indirectly, which is different from traditional face-to-face communication. It is common interests that make people who do not know each other work together to create, modify, and perfect entries.

**Analysis on Interest of Contributors**

It is found that entries are connected with experts or users as some entries have the common contributors. Coupling analysis is used to reveal the relationship of entries which are results of common experts or users, which could also indicates the common interest of contributors.

**Common Interest of Experts**

Entries coupling analysis based on experts is shown as Fig.5. There are many clusters of entries. It is hard to master all of branches of Biomedical Science for every expert. It is said that specialist only masters his own field. Entries coupling analysis based on experts will show the relevance of these entries to some extent though the entries seemed isolated. For example, some experts (e.g. Gu Hanqing) who focused on Tissue Engineering and artificial organs contributed to the entries including “artificial skull”, “artificial skin”, “artificial kidney”, and so forth. From the related entries clustered together, branch of discipline and research hotspot could be indicated.
Common Interest of Users

Entries coupling analysis based on users is shown as Fig.6. It is a huge and complex network which is different from the network based on experts. The entry “contact lenses” is the contribution of 236 users, whose number of users is the most. There are 1379 users who contributed to the entries. It is found that the users were of multiple interests. The user “csbme” contributed the most entries whose number was up to 237. There are at least 83 users who contributed to no less than two entries. For example, the user “yangke19941112” contributed to 11 entries while “w_ou” contributed to 47 ones. It is found that the entries were collected together because of the active contribution of the users. It seemed that the users were interested in many entries in the field of Biomedical Science.

The co-contributed entries can be considered as medium which is of great importance during the communication among contributors. Each contributor expressed one’s ideas in the form
of modifying content of entries. Then others could see, think and express their own opinions by further modification. The common interested entries became the platform on which contributors expressed their views. The modification would not end until the contributors and readers were satisfied with the content. The whole process contains a number of communication activities.

**Conclusion and Discussion**

In general, there are four questions answered in the paper. The composition of related entries on Baidu Baike is analyzed with citations among entries. It is found that citations among entries involved in other disciplines beyond Biomedical Science. It is not a simple citation activity but a kind of signal of disciplinary crossing and integration, which is the main trend of modern science and technology development. The crossing and integration of disciplines is not only growth point of new disciplines, but also leads to the emergence of scientific discoveries of great importance.

The relationships between online entries and traditional literatures have been discussed based on analysis of cited references of each entry. It shows that traditional knowledge carriers such as books, journals, conference have still profoundly affect people’s communication habits. These traditional knowledge carriers are considered reliable, scientific, and reasonable.

Cooperation network is drawn to show how experts or users work together. It seems that majority of experts prefer to cooperate with someone who works in the common affiliation while cross-unit cooperation also exists. For example, experts from Peking University and Tianjin Medical University cooperated on the entry “cochlear implant technology” and “Artificial dura mater”. The experts from the Hong Kong Polytechnic University and Suzhou University collaborated on entries related to ankle movement and its mechanical function.

And coupling analysis of entries is used to reveal the common interest of contributors. The clusters of coupling entries based on experts indicate that the specialist only masters his own field while users have wide range of interests and are fired with an enthusiasm for entry contribution. The development of science does not only depend on scientists or government, but also need participation of the general public. Encouragement to users to devote to entry contribution should be attached great importance to.

However, there are still some questions which need to be answered in the future research.

*How did experts collaborate or cooperate with users during the contribution of the entries?*

According to the historic versions of each entry, the modified time and reason of users could be found out. But we could hardly find out when and how the experts interfere with the content of entries. As a result, it is very difficult to know exactly the way experts cooperate or communicate with users. Maybe the questionnaires can be conducted on the users and experts in order to investigate the process.

*What factors did affect the entry contribution? What are essential factors on communication among the contributors?*

To answer these questions, personal interviews and questionnaires should be conducted on the base of quantitative analysis which is of great help to find out the target objects. It is necessary and vital to adopt combined quantitative analysis and qualitative analysis.

*How about the entries of other fields or other disciplines beside of Biomedical Science? How about other online encyclopedia?*

The number of studied entries should be increased as 247 entries are only a drop of the ocean compared with all the entries on online encyclopedia. And the Biomedical Science entries
were analyzed. The preliminary work of data collection was conducted manually. It is necessary to develop a web crawler to pull the data from each entry. It is known that a new trend of interdisciplinary research is increasingly common. The citation among entries from different disciplines can be expected. It is necessary to collect enough data from different disciplines to reveal interdisciplinary integration behind potential relationship of the entries.

**Acknowledgments**

The paper is supported by National Social Science Foundation of China (Grant No. 15CTQ024) and the National Natural Science Foundation of China (Grant No. 71603195).

**References**


Are Scientometrics, Informetrics, and Bibliometrics different?

Yang Siluo\textsuperscript{1} and Yuan Qingli\textsuperscript{2}

\textsuperscript{1}58605025@qq.com
Wuhan University, Wuhan (China)

\textsuperscript{2}331061947@qq.com
Wuhan University, Wuhan (China)

Abstract
Bibliometrics, scientometrics, and informetrics (also called the three metrics) differ in subject background but are the same in theories, methods, technologies, and applications. Analyzing their current situation and relationships can help comprehensively understand the three fields. In this study, we collect the data of the three metrics through keyword search in 2007–2016. We also compare and visualize the three metrics in terms of the distribution of publications and cooperation (recognition level), the main research topics (intellectual structure), and the reference situation (knowledge communication). Results show that the three metrics differ in the degrees of utilization and recognition but are similar in the general direction. We recommend the addition of bibliometrics in the title of the International Society for Scientometrics and Informetrics.

Keywords
Bibliometrics; Scientometrics; Informetrics; Knowledge Domain Visualization; CiteSpace

Conference Topic
The relationship and development of five metric science concepts: Bibliometrics, Informetrics, Scientometrics, Webometrics, and Knowledgometrics.

Introduction
Bibliometrics, scientometrics, and informetrics (also called the three metrics) are three related terms in metrology. These terms are used to describe similar and overlapping methodologies; however, their well-documented historical origins differ, and they are not necessarily synonymous (Hood & Wilson, 2001). The rapid progress and development of science and technology have improved the research objects, goals, and methods of the three metrics. New branches such as webmetrics and altmetrics (Egghe, 2005), and new indexes and evaluation measures including Citescore and the H index, have also appeared (Hirsch, 2005). Although these terms differ in disciplinary background and emergence time (Qiu et al., 2017), they are used in accordance with their own cognition and position, thereby causing significant confusion. There are many journals with “scientometrics” or “informetrics” in their titles exist, but few journals with “bibliometrics” in their titles exist. The International Society for Scientometrics and Informetrics (ISSI) is the most significant conference in the three metrics. ISSI contains scientometrics and informetrics in its title but has no bibliometrics. Given that confusion can harm the development and application of the three metrics (Hood & Wilson, 2001), the development of the three terms and their relationships should be examined. This study explores the current situation and relationships of the three metrics from the following three aspects: the number of published papers and cooperation (recognition level), the main research topics (intellectual structure), and the reference situation (knowledge communication).
Background

Definitions of bibliometrics
Bibliometrics methods have been applied in various forms for more than a century (Hood & Wilson, 2001). The term “bibliometrics” was first introduced by Pritchard (1969), who defined it as “the application of mathematical and statistical methods to books and other media of communication.” Fairthorne (1969) expanded the definition scope of bibliometrics by defining it as the quantitative treatment of the properties of recorded discourse and behavior appertaining to it. Then, Broadus (1987) defined bibliometrics as the quantitative study of physically published units, or of bibliographic units, or of alternatives of either.

Definitions of scientometrics
Nalimov and Mulchenko (1971) coined the Russian equivalent of the term “scientometrics” in 1969, and defined it as the quantitative study of various kinds of intelligence process in the development of science. The term has obtained broad acceptance from the journal Scientometrics, which was built in 1978. Scientometrics is a discipline that uses mathematical methods to quantify the scientific research personnel and achievements to reveal the process of scientific development, and can provide scientific basis for scientific decision making and management (Qiu et al., 2017). Scientometrics uses citation analysis and other quantitative methods to evaluate scientific research activities and thus guide the policy of science (Egghe, 2005).

Definitions of informetrics
Nacke first proposed the German term “informetrie.” By the early 1990s, the term “informetrics” obtained wide recognition. Nacke believed that informetrics is a study applied in mathematical methods for information science objects (Qiu et al., 2017). This definition is slightly one sided because it limits the scope of informetrics in information science. Tague-Sutcliffe extended informetrics to the quantitative study of any form of information; thus, informetrics is not simply a bibliographic record or any social group, or not limited to scientists (Fairthorne, 1969). This definition enlarges the research scope and content of informetrics. Qiu et al. (2017) divided informetrics into two aspects of broad and narrow senses. The broad sense of informetrics research is very broad, whereas the narrow sense of informetrics mainly uses mathematical, statistical, and other quantitative methods to study the characteristics and laws of information quantitatively.

Relationships among the three metrics
The three terms have evolved to share many of the objectives and have many methods and tools in common (Qiu et al., 2017). The three metrics refer to “component fields related to the study of the dynamics of disciplines as reflected in the production of their literature” (Hood & Wilson, 2001). The three terms often appear simultaneously, or used interchangeably by authors, such as the Second International Conference on Bibliometrics, Scientometrics, and Informetrics (now called “ISSI”). However, the terms differ in their discipline attribute; specifically, bibliometrics belongs to library and document science, scientometrics belongs to the science of science, and informetrics belongs to information science (Brookes, 1990; Qiu et al., 2017; Wang, 1998). Scientometrics and informetrics have been proposed for nearly 50 and 40 years, respectively; however, they lack their own uniform concepts that can be widely accepted by the public. Different definitions of bibliometrics also exist (Hood & Wilson, 2001). The relationship among the three metrics has long been investigated. Brooks (1990) explored the origin and interrelationship of the three metrics. Glänzel and Schoepflin (1994)
emphasized that the authors’ use of “bibliometrics” synonymously for the three metrics has resulted in chaos. Hood and Wilson (2001) analyzed the differences among the three metrics by investigating the history of the three terms through analyzing the number of papers and journals between 1968 and 2000. Wen and Qiu (2006) suggested that the three metrics belong to different superordinate disciplines; however, they have the same research objects, indicators, and methods. Some believed that the three metrics present a crossing and partial overlapping relationship, but others argued that the three metrics exhibit an inclusive relationship; for example, informetrics has many meanings and includes bibliometrics and scientometrics (Qiu et al., 2017).

Data and method
A subject or discipline is often analyzed using the literature statistics, which accesses data samples in two ways: (1) choosing the top journals or core journals in the field as the data source (Milojevic & Leydesdorff, 2013), and (2) obtaining data source through the retrieval of representative keywords (Hood & Wilson, 2001). The journals in the three metrics present great repeatability and are widely distributed. Thus, this study uses keyword research to obtain data for the comprehensive comparison of the differences among the three metrics.

Method
The current situation and relationships of the three metrics include various aspects. In this study, we analyze only three major areas: 1) the usage and distribution of the three terms on the basis of the number of publications and cooperation; 2) the research contents and intellectual structure on the basis of field topics; and 3) the knowledge communication and flow on the basis of the citation and reference. We use EXCEL and CiteSpace for data statistics and network development.

Data
We downloaded data sets comprising articles, reviews, and papers from SCI-Expanded, SSCI, and A&HCI between 2007 and 2016. Following Hood and Wilson (2001), we retrieve the literature of the three metrics using “TS = (bibliometric* or bibliometry or bibliometrical* or bibliometrician* or “statistical bibliography” or bibliometrie),” “TS = (scientometric* or scientometry or scientometrical* or scientometrician*),” and “TS = (informetric* or informetry or informetrician* or informetrie).” Finally, we combine the three search strategies to investigate the overlapping situation of the three metrics.

Result

Overall situation
The retrieval results show that bibliometrics has the largest number of publications, which is approximately four times of that of scientometrics and 20 times of that of informetrics. A large difference is found in the number of literature among the three fields in 2007–2016, possibly because of the difference in their history and degree of social recognition. In the past 10 years, the annual volume of publications on bibliometrics is higher than that on scientometrics and informetrics, the number of publications on bibliometrics has the largest increase among the three fields, the increase for scientometrics is at an intermediate level and informetrics presents the smallest increase.
**Figure 1** Search results of combining the three metrics

**Distribution of publications and cooperation**

**Number of publications in different levels**
The European countries dominate more than half of the top 10 national rankings of the three metrics, indicating that Europe is the core area of international research on the three metrics. Europe is the origin of the three fields and has a long history of research; thus, it is home to popular research institutions and experts. With regard to American countries, the United States and Brazil as the main country, the number of publications in the field of bibliometrics and scientometrics are higher than informetrics. Scientometrics has been widely explored and is highly recognized and utilized in Asia. In Oceania areas, the degrees of utilization and recognition of bibliometrics are much higher than those of scientometrics and informetrics. Regarding African countries, the number of articles on the three areas is few and the level of scientific research and production is low.

At the institutional level, universities account for the majority of the top 10 institutional rankings of the three metrics. A total of 6, 7, and 4 European institutions enter the top 10 institutional rankings of bibliometrics, scientometrics, and informetrics. This finding shows that European institutions play an important role in the three metrics. The main institutions of Asia are mostly from China. With regard to American countries, only Indiana Univ enters the top 10 institutional rankings of bibliometrics and scientometrics, and also ranks at the top in informetrics. In African countries, South Africa’s Univ S Africa ranks fourth in informetrics. This finding combined with the number of articles of South Africa shows that South Africa presents strong research strength in informetrics.

At the author level, the top 10 high-yield authors in bibliometrics mainly include Bornmann, Abramo, and Ho; those in scientometrics mainly include Groneberg, Ho, and Leydesdorff; those in informetrics mainly include Egghe, Rousseau, and Burrell. Egghe and Rousseau enter only the top 10 author rankings in informetrics.

**Cooperation network**

Author cooperation network

Figure 2 shows that seven large groups are present in the author cooperation network of bibliometrics in 2007–2016. Notably, the size of the nodes represents the number of publications of authors. D’Angelo, Abramo, Bornmann, and Ho, and other high-yield authors, have their own fixed partners and present close cooperation within the group. Some authors publish few articles but present many partnerships. For example, Waltman is in Ho’s collaboration group and has a small number of articles but has six cooperative partners in the threshold range. Moreover, Kostoff publishes few articles but collaborates with every member.
of the group. The said authors are inclined to conduct scientific research through cooperation. In general, many cooperative groups exist in bibliometrics. The internal cooperative relationship is close, collaborations between only two authors are few, and the degree of cooperation is high.

Figure 2 Author cooperation network of bibliometrics
Figure 3 shows that scientometrics presents a large cooperation group and that Groneberg is the center of the group. Leydesdorff and Bornmann are included in this group; however, the two authors locate in the extension of the group. This group exhibits an intricate connection and a close relationship. A prominent cooperative group exists, in which Ho of four fixed partners is the center. Therefore, this group has a stable partnership. Many two- and three-author groups exist. In general, the scale of the author collaboration network of scientometrics is sparse.

Figure 3 Author cooperation network of scientometrics
Figure 4 shows that informetrics has a main cooperative group, in which Egghe and Rousseau are the core, and includes Ye and Liu, among others. Egghe and Rousseau publish the largest number of articles on informetrics. The rest of the groups is dispersed. Apart from the largest group, more than 10 three- and two-author groups exist. The scale of the author cooperation network is small.
The size of author cooperation depends on two main factors: the total number of articles and the habit of research. With regard to the first factor, authors with large numbers of articles have high numbers of cooperative relationships, thereby leading to large-scale cooperation. For the second factor, some authors prefer to cooperate with other authors, thereby forming large-scale and stable cooperation. Under the same thresholds, the number of large-scale cooperative groups (more than five authors) in bibliometrics is more than that in scientometrics and informetrics; however, the group relationships of the three metrics are tight. Bibliometrics obtains the maximum number of literature and degrees of utilization and recognition. The cooperative group with Groneberg as its core in scientometrics is the largest among the three fields. The cooperative relationships among members are also frequent in the three metrics. The decentralization degree of the author collaboration network of informetrics is the sparsest among the three fields. Although its size of cooperative groups is the smallest, informetrics possesses a major group with Egghe and Rousseau as its core.

Institutional cooperation network

In bibliometrics, the connection among institutions is tight and cooperation between institutions is close. Europe is home to many research institutions, which attach great importance to cooperation in bibliometrics. Only the Chinese Acad Sci has a purple aperture in the cooperative network. Therefore, this institution presents high centrality and is an important mediator in the entire cooperative group to date. Leiden Univ, Univ Granada, Asia Univ, and CSIC exhibit many cooperative relations and publish several articles on bibliometrics. Some institutions are very successful in cooperation even if they do not publish many articles, such as Univ Amsterdam and Univ Carlos III Madrid.

The institutional collaboration network of scientometrics shows three large cooperative groups (more than five institutions). In the first group, Asia Univ and Univ Amsterdam and other high-yield organizations are the core. This group presents the largest number of cooperative institutions among other groups, but its internal cooperation is sparse. In the second group, Goethe Univ Frankfurt is the core. In the third and final group, Univ Granada and CSIC are the core. In general, scientometrics presents a large number of cooperative groups but in a small scale.

During the study period, informetrics possesses a significantly large cooperative group with Katholieke Univ Leuven and Univ Antwerp as the core. This group also includes Univ Hasselt and KHBO Assoc KU Leuven and other institutions. Notably, European institutions account for the majority of the group. Some Chinese institutions are also present in this group, such as Zhejiang Univ, Chinese Acad Sci, and Nanjing Univ. Therefore, Chinese and European institutions exhibit frequent cooperation in informetrics. Indiana Univ (American) and Dalian Univ Technol (Chinese) assume the role of intermediaries in their respective groups. Although they are excluded in the main cooperative group, their degree of cooperation is high.
In general, under the same parameter setting, the degree of cooperation among the institutions of bibliometrics is the highest, and the cooperative relationship of bibliometrics is close and of the largest scale among the three metrics. Meanwhile, the cooperative relationship of scientometrics and informetrics is dispersed. Only one to three large groups exist in scientometrics and informetrics; thus, the scale of other groups is very small. However, the scale of institutional cooperation is also affected by the total amount of literature and the habit of specific authors.

**Subject structure**

We choose the co-words analysis to explore the subject structure of the three fields. Considering the color confusion after clustering, we provide a screenshot of categories depending on the color of each node prior to labeling the category name. The latter analysis is based on the cluster name provided by Citespace. The clustering results are shown in Figs. 5, 6, and 7.

Figure 5 shows five main subjects in bibliometrics: 1) research on the general development trend and influence of bibliometric analysis; 2) research based on bibliometric indicators of the scientific research output, the ranking of universities, and the evaluation on individual academic; 3) application of bibliometrics in scientific research management; 4) research on cooperation network and model, application of text mining, and other new technologies in other disciplines; 5) the use of the H index and other indicators to analyze the citation of papers or other publications from various databases and thus evaluate the academic capability and scientific research achievements.

Figure 6 shows five main subjects in scientometrics; 1) application of scientometrics methods in biology, indicating the extensive application of scientometrics methods in other fields; 2) research on the quality of scientific research output and the research trend of scientometrics as a branch of science; 3) research on scientometrics methods (such as citation analysis and impact factors) based on journals and other research outputs, indicators (such as the H index), and their application; 4) development trend of scientific cooperation, cooperation model, and academic cooperation network based on scientific research output; 5) ranking or visualization of discipline contents, personnel, and journals through scientometrics methods, as well as research on the development trend of disciplines.
Figure 6 Subject structure of scientometrics

Figure 7 shows five main subjects in informetrics; 1) research and application of the H index and other new metrics in network environment, as well as research on information systems; 2) research on the influence of articles and authors on the basis of informetrics methods and citation analysis; 3) research on the distribution and ranking of high-impact articles, authors, and journals based on the H index, as well as research on the model of informetrics; 4) research on the development of informetrics and its relationship with bibliometrics and scientometrics; 5) effect of network environment on the patterns of scientific research activities.

Figure 7 Subject structure of informetrics

According to the keyword frequency, “science,” “citation,” “impact,” “Journal,” “citation analysis,” “H index,” and “impact factor” rank among the top 10 keywords in the three areas. Therefore, the three fields are concerned on the research and evaluation of the influence of scientific research output. Most studies use the citation analysis method and pay attention to the innovation and application of indexes.

In summary, the three metrics focus on evaluating scientific research output and investigating the innovation and application of indexes (such as the H index) and the cooperation mode and network. Bibliometrics focuses on exploring the development trend and research on the application of bibliometrics in scientific research management. Scientometrics focuses on exploring the application of its methods and techniques in other areas, the development of citation analysis and other methods, and the quality of scientific research output and development trends. Given that information is highly dependent on computer technology and
mathematics, informetrics focuses on examining information system and the combination of mathematical models and methods.

Citation situation—Mainly based on document co-citation network
According to the document co-citation network, the knowledge structure of bibliometrics includes five components. Scientometrics is composed of six knowledge bases. Given that three of these bases have the same theme (but different direction), we merge the three identical parts into one. Finally, four components are obtained. The knowledge base of informetrics consists of five parts. The connection between nodes in the graph represents the citation relationship between the documents; the node with the purple aperture represents its high important, and the size of the node represents the cited frequency. We compare the three metrics on the basis of the cited literature with high centrality and high cited frequency. According to the specific literature list, Hirsch’s *An index to quantify an individual’s scientific research output* and Egghe’s *Theory and practice of the G index* are important documents in the three fields. The two articles have been used as the bases for the study on the H index in bibliometrics and scientometrics. Meanwhile, Hirsch’s article has also been used as the basis for the study on evaluation and ranking in informetrics. Therefore, the same literature is used in different angles in dissimilar fields. Hirsch (2005) proposed the H index to evaluate the academic influence and the research level of researchers; this document is the first of the branch. Egghe (2006) proposed the G index in 2006 to overcome the shortage of the H index, and used the former index to measure the overall citation performance of a group of articles. The H index is the common knowledge base of the three areas, which indicate that the three metrics have attached importance to research on the innovation and application of evaluation indicators.

As shown in Figure 7 and Table 1, #0 research trend, #3 scopus, and #4 gender disparity are peculiar knowledge components of bibliometrics. #0 research trend involves the review on the history of each specific research direction and the future prediction under the premise of innovation of indicators and progress of research methods. Alonso et al. (2009) studied the H index and the subsequent derivative index, and measured the application of the H index in different fields. Bornmann and Daniel (2008) conducted a review of scientists’ citation behavior and analyzed the motivations, impact factors, and trends of the citation. #3 scopus presents diverse data sources in the context of the development of network; in this part, researchers focus on emerging data search tools and their data coverage, search capabilities, and the impact of scientific research (Meho & Yang, 2007; Falagas et al., 2008). #4 gender disparity is the part wherein researchers in medicine evaluate the academic productivity and influence of researchers using the H index to study whether the gender differences can lead to differences in academic productivity (Eloy et al., 2012).

![Figure 7 Literature citation figure of bibliometrics](image-url)
As shown in the document co-citation network and the specific literature list of scientometrics, #1–4 scientometric approach and #5 scientometric are peculiar knowledge components of scientometrics. #1–4 scientometric approach is the basis of two research directions. King measured (2004) the status and influence of national scientific research with scientometrics methods, and Moed (2005) explored the application of citation analysis in scientific evaluation. The two documents provided an empirical basis for the application of scientometrics methods in the scientific evaluation. Konur applied scientometrics methods to the quantitative research in biochemistry, and this document provided the knowledge base for the application of scientometrics methods to different fields. #5 scientometric mainly contains the scientific basis of visualization technology and social network analysis in scientometrics. Chen (2006, 2010) developed CiteSpace, which is an important visualization tool for studies on subject topics and trends, and its functions are constantly optimized and improved to date.

As shown in the document co-citation network and the specific literature list of informetrics, #0 ranking, #2 pagerank, and #4 Lotkas law are peculiar knowledge components of informetrics. #0 ranking is mainly based on the articles of Hirsch (2005) and Braun et al. (2006), who studied the assessment and ranking of academic impact on individuals and journals using the H index, thereby establishing the foundation for the innovation and application of indicators in ranking and evaluation of scientific research. #2 pagerank is the part wherein Ding (2009) studied factors that affect the ranking of the PageRank algorithm, and proposed the weighted PageRank algorithm. The study provided experience for the further application of the PageRank algorithm in informetrics. #4 Lotkas law is the part wherein the basic knowledge theory is comprehensively explored and the mathematical model is applied to the theoretical hypothesis.

## Discussion and Conclusion

We retrieve the literature (such as articles, reviews, and proceeding papers) on the three metrics from SCI-Expanded, SSCI, and A&HCI using the three terms between 2007 and 2016. The results show that the three metrics differ in the degrees of utilization and recognition but are similar in the general direction.

1) Recognition. Combining the number of articles in national and institution levels shows that Europe is the core area of the three fields. The degrees of recognition of bibliometrics and scientometrics are high in America. The degree of recognition of scientometrics is higher than that of the two other fields in Asia, and the degree of recognition of bibliometrics is much higher than that of the two other fields in Oceania. The degrees of recognition of the three fields are low in Africa, but the degree of recognition of informetrics is high in South Africa. Bibliometrics is the most frequently used and has the largest degree of increase among the
three terms, so we think that the term "bibliometrics" can be used as a general term for "scientometrics" and "informetrics" in order to avoid confusion in terms.

2) Cooperation. Universities are the main institutions in the three metrics. The scale of the cooperative group of bibliometrics is larger than that of scientometrics and informetrics, and the group relationships of bibliometrics are tight. Many two-author cooperative groups with scattered scale exist in scientometrics. The author cooperative network of informetrics is the sparsest and has the smallest scale among the three metrics. Bibliometrics has the highest degree of institutional cooperation and the largest scale of institutional cooperation network. On the contrary, the cooperative relationships of scientometrics and informetrics are dispersed, and the size of the group is small.

3) Subject structure. Bibliometrics attaches great importance to the development of itself and the application of bibliometric methods in scientific research management. Other disciplines use scientometrics when applying metrological methods, such as citation analysis. Scientometrics emphasizes the quality of scientific research output and focuses on research on scientometrics development trends. In the network environment, researchers prefer to use informetrics when researching information systems and the combination of mathematical models and informetrics methods.

4) Knowledge base. A significant difference is found in the knowledge base of the three metrics. The H index is the common knowledge base of the three fields in focusing on innovating and improving indicators. The peculiar knowledge bases of bibliometrics include review on the history of each specific research direction, data coverage and search capabilities of emerging data search tools, and application of bibliometrics in the medical field. The peculiar knowledge bases of scientometrics include the application of scientometrics methods in scientific evaluation and other disciplines, such as visualization techniques and social network analysis. The peculiar knowledge bases of informetrics include comprehensive exploration of the basic theory of and research on the PageRank algorithm.

This study presents a few limitations. The overall situation of the three metrics can be accurately compared using long data period. However, this study uses only the data from 2007 to 2016. Although we carefully refine some terms and construct a document retrieval formula, the data collection is still incomplete and thus cannot fully represent the data for the three metrics. The current situation and relationships of the three metrics include various aspects, but this study focuses on only three major areas. Therefore, the comparative analysis is insufficient. Future research can conduct interview with experts, extend the data period, perform in-depth content analysis, and use other methods in comparing the relationships among the three metrics.

Acknowledgments
This research is funded by A Foundation for the Author of National Excellent Doctoral Dissertation of PR China (2014094).

References


Webometrics of Think Tanks in Sub-Saharan Africa

Omwoyo Bosire Onyancha¹ and Williams E Nwagwu ²

¹ onyanob@unisa.ac.za
University of South Africa (Pretoria, South Africa)

² williams.nwagwu@codesria.sn
CODICE, Council for the Development of Social Science Research in Africa (Dakar, Senegal); Research fellow: Department of Information Science, University of South Africa (Pretoria, South Africa)

Abstract
This paper reports preliminary findings of an ongoing project on the assessment of the performance of think tanks (TTs) in sub-Saharan Africa. The purpose of this paper was to investigate the performance of the TTs on the web through the webometrics analysis of their websites. In crawling the websites of the TTs, using the SocSciBot crawler, the study found that a sizable number of TTs in sub-Saharan Africa do not have their own websites; TTs in South Africa performed much better in terms of web pages and web links than the TTs in the rest of sub-Saharan Africa; South African TTs provide compact interlinkages as opposed to those in other sub-Saharan African countries; although there were positive correlations between in- and out-links, the relationship was not significant; the external links largely targeted organizations (.org) and commercial companies (.com); subjects or topics reflected in the external links included business, women, agriculture, politics, law, youth, climate, economy, climate change, development, growth, jobs, peace, budget, democracy, partnership, security. The study has major implications as far as the web presence, reach and impact of TTs as well as their performance and accountability are concerned.

Conference topic
Think tanks construction and evaluation

Introduction
The role and value of think tanks (TTs) in socio-economic development as well as shaping policies in political and governance arenas in developing economies such as those in sub-Saharan Africa cannot be overemphasised (Abelson 2000; Johnson 2009; Mbabazi, MacLean & Shaw 2002). Mbabazi, MacLean and Shaw (2002) have observed that TTs have emerged in Africa as major producers of the literature on the political economy of violence. The authors argue that, in by so doing, the TTs have helped to explain the link between African conflict to the “imperatives of production and consumption in relation that juxtaposes Africa’s political institutions and cultures with international and global economies”. The role of TTs is more pronounced in linking knowledge to government policy or policies effecting civil societies in which the TTs are located. McGann (2015: 6) puts it thus: “across both developed and developing countries, governments and individual policymakers face the common problem of bridging expert knowledge to bear on government decision-making. Policymakers need reliable, accessible, and useful information about the societies they govern.” Simply put, TTs exist solely to support self-sustaining economic, social and political progress of all regions of the world for public good (McGann 2015). There has been a rapid growth in the number of TTs in the world, while the scope and impact of TTs’ work has equally expanded, just as has their potential to support and sustain democratic and civil societies (Krastev 2008; McGann 2015: 9). In Africa, TTs are viewed as agents of shaping the African futures because “they can be used to influence and inform public policy in real time and space” (Mbadlanyana, Cilliers & Sibalukhulu 2011: 65). TTs draw their strength from their ability to act as a bridge between the academic and policymaking communities (Mbadlanyana, Cilliers & Sibalukhulu 2011: 65).
Yet, as recently as 2000, Stone (2000) noted that policy research institutions and TTs were a neglected phenomenon in the social science literature. TTs in Sub-Saharan African are said to be challenged on several fronts, some of which are weak financial base and dependence on donors (especially foreign ones) and human resource challenges, which include inadequate policy analytical expertise in the staff, too few people with management expertise, stiff competition for a small pool of competent and committed staff, lack of opportunities for training core staff in areas of operations, high staff turnover, and a lack of international exposure among staff (Johnson 2009: 486). Johnson (2009: 486) further explains that TTs in sub-Saharan Africa suffered from a dilapidated and an underdeveloped state of infrastructure, especially in the area of telecommunications. Johnson (2009) noted that, by 2009, African think tanks not only were difficult to reach by phone, they also had very limited access to or presence on the Internet … only two groups outside of South Africa had their own websites by then. Mouton (2010), too, has reiterated the absence of sub-Saharan Africa’s research institutes, centres and networks on the World Wide Web. By the time of publishing his findings in the UNESCO report in 2010, Mouton (2010) found that only 53% of the institutions had active websites, but hastens to clarify that “active” websites did not necessarily mean that the content on the websites was current. Mouton (2010) further has vividly captured the unfortunate state of research institutions (including TTs) in sub-Saharan Africa by observing that the institutions hardly receive any government support, with the exception being the Human Sciences Research Council (HSRC) in South Africa.

### Problem statement and purpose of study

The problem statement for the current study is threefold. Firstly, we have noted that the Webometrics Ranking of Research Centres (WRRCs) (Cybermetrics Lab 2015) does not cover all TTs in sub-Saharan Africa, most probably due to the nature of some of the TTs which may not be considered as research centres. Secondly, the Think Tanks and Civil Societies Program (TTCSP) at the University of Pennsylvania ranks TTs according to a number of variables such as influence, output and impact. The ranking system does not consider one of the variables that have increasingly become valuable in assessing the performance of institutions, namely the web-based indicators, for example, web presence, visibility and impact (Aguillo 2009; Aguillo, Ortega, Fernandez & Utrilla 2010; Kenekayoro, Buckley & Thelwall 2014). It is widely acknowledged that the Web has increasingly become one of the essential tools with which not only organizations market themselves but communicate their research and share their policy documents thereby influencing policy changes in their respective countries or geographical context (Thelwall, Klitkou, Verbeek, Stuart, & Vincent 2010). Finally, sub-Saharan Africa is home to a large number of TTs, yet it is one of the least investigated regions by scholars in terms of the TTs’ performance on the Web.

### Purpose of the study

The purpose of this study was therefore to assess the Web visibility and reach of think tanks in sub-Saharan Africa through webometrics means. The study investigated the number of pages, outlinks and inlinks as well as the content of the links and pages in order to gauge the performance of the TTs in sub-Saharan Africa on the Web.

### Methods and materials

The study adopted a webometrics approach to investigate the presence, visibility and impact of think tanks (TTs) in sub-Saharan Africa on the World Wide Web, simply and hereafter referred to as the Web. A list of the TTs, as provided in Appendix A, was obtained from the 2014 Global Go to Think Tank Index Report, prepared by James McGann (McGann 2015: 1520).
The list contained a total of 91 TTs for sub-Saharan Africa. All TTs were included in this study for investigation. Once the TTs were identified, a Google search for their official websites was conducted to obtain the URLs (Uniform Resource Locator), or simply put, the web addresses which were in turn used to crawl the web for data. A total of 66 TTs were eventually selected for the study as these are the only ones that owned websites. The SocSciBot web crawler was used to obtain data from the TTs’ websites. According to the developer of SocSciBot, Mike Thelwall, the “SocSciBot, like any web crawler, works by recursively requesting web pages. It extracts URLs from the HTML of web pages and repeats the process with each new URL found” but is limited in that it does not “check for pages duplicated across different sites” (Thelwall 2015a). As Thelwall (2015b) observes, the limitation only becomes an “issue if a site crawled contains a large mirror site” (Thelwall 2015b: 14). The search for each of the TT’s web data was conducted in the months of July to September 2016. The search limits were set as follows in order to extract relevant data: (a) the number of URLs to crawl was set to a maximum of one million; and (b) the URLs with question marks were omitted.

Data analysis was conducted using the analyse links with SocSciBot tools option that is embedded in the crawler. The analysis generated the following files that contained pertinent data for the current study: ADM count summary (which provided standard link counting data); all external links (a list of URLs targeting pages outside of each site, i.e. site outlinks); all links with counts (a list of all links that match the current project options and accompanied with link counts); directory interlinking (the number of links, counted directory by directory, i.e. the number of pairs of directories with a link from the first to the second directory [directory ADM]); domain interlinking (the number of links, counted domain by domain, i.e. the number of pairs of domains with a link from the first to the second domain [domain ADM]); file interlinking (the number of links, counted page by page, i.e. the number of pairs of pages with a link from the first to the second page); page and link counts (the number of pages in each crawl; and the number of links found in the web pages of each crawl).

In addition, the crawler provided some information that was used to discuss the findings obtained in the above outlined files. This additional information included:

- Selected external links with counts (a list of links, with counts, from each site to the other sites crawled, excluding self-links)
- Unselected external links with counts (a list of all links, with counts, from each site to other sites except the crawled sites).

The analysis tool also produced a map of the file and domain interlinkages among the TTs. The map was prepared and produced using the Pajek software. Finally, the most common words within the links were analyzed using the VosViewer software and visualized on a map to reveal the content of the web linkages amongst the TTs.

**Results and discussions**

**Number of pages and outlinks**

Collectively, the TTs that were under investigation in this study produced 3.3 million pages, which have had over 140 million outlinks by September 2016. The TTs that yielded over 500 thousand pages included the *African Centre for the Constructive Resolution of Disputes* (ACCORD), which yielded approximately one million pages that accounted for 30% of all the number of pages of TTs in sub-Saharan Africa; *South African Institute of Race Relations*
[IRR] (915 213 pages, 27.58%); Economic Policy Research Centre [EPRCU] (740 454 pages, 22.32%); and Afro-Middle East Centre [AMEC] (514 486 pages, 15.51%). In terms of the number of outlinks, the four institutions yielded the most. Collectively, they had 136 million outlinks, which constitutes approximately 97.16% of all outlinks. Among the top TTs in terms of the number of pages, the Free Market Foundation [FMF] yielded the highest number of outlinks per page (i.e. 72.47) followed by the Ethiopian Economics Association (62.19), AMEC (60.59), EPRCU (57.49) and the United Nations Economic Commission for Africa [UNECA] (56.31). The analysis of the correlation between the number of web pages and outlinks yielded a positive Pearson correlation coefficient (r) of 0.926 (where N = 66), which was significant at p \leq 0.01.

**Table 1: Number of pages and links per TT in sub-Saharan Africa**

<table>
<thead>
<tr>
<th>Site</th>
<th>Pages</th>
<th>% pages (N=3317899)</th>
<th>Outlinks</th>
<th>% outlinks (N=140023850)</th>
<th>Outlinks/page</th>
</tr>
</thead>
<tbody>
<tr>
<td>irr.org.za</td>
<td>915213</td>
<td>27.58</td>
<td>41133444</td>
<td>29.38</td>
<td>44.94</td>
</tr>
<tr>
<td><a href="http://www.eprcug.org">www.eprcug.org</a></td>
<td>740454</td>
<td>22.32</td>
<td>42572080</td>
<td>30.40</td>
<td>57.49</td>
</tr>
<tr>
<td>amec.org.za</td>
<td>514486</td>
<td>15.51</td>
<td>31170862</td>
<td>22.26</td>
<td>60.59</td>
</tr>
<tr>
<td><a href="http://www.saiia.org.za">www.saiia.org.za</a></td>
<td>33498</td>
<td>1.01</td>
<td>50241</td>
<td>0.04</td>
<td>1.50</td>
</tr>
<tr>
<td><a href="http://www.ccr.org.za">www.ccr.org.za</a></td>
<td>23996</td>
<td>0.72</td>
<td>909511</td>
<td>0.65</td>
<td>37.90</td>
</tr>
<tr>
<td><a href="http://www.uneca.org">www.uneca.org</a></td>
<td>19859</td>
<td>0.60</td>
<td>1118278</td>
<td>0.80</td>
<td>56.31</td>
</tr>
<tr>
<td>aercafrica.org</td>
<td>11957</td>
<td>0.36</td>
<td>500157</td>
<td>0.36</td>
<td>41.83</td>
</tr>
<tr>
<td><a href="http://www.mefmi.org">www.mefmi.org</a></td>
<td>8254</td>
<td>0.25</td>
<td>63817</td>
<td>0.05</td>
<td>7.73</td>
</tr>
<tr>
<td><a href="http://www.fanpan.org">www.fanpan.org</a></td>
<td>8241</td>
<td>0.25</td>
<td>258947</td>
<td>0.18</td>
<td>31.42</td>
</tr>
<tr>
<td><a href="http://www.ossrea.net">www.ossrea.net</a></td>
<td>5764</td>
<td>0.17</td>
<td>42137</td>
<td>0.03</td>
<td>7.31</td>
</tr>
<tr>
<td><a href="http://www.igd.org.za">www.igd.org.za</a></td>
<td>4852</td>
<td>0.15</td>
<td>68329</td>
<td>0.05</td>
<td>14.08</td>
</tr>
<tr>
<td><a href="http://www.acode-u.org">www.acode-u.org</a></td>
<td>3551</td>
<td>0.11</td>
<td>121987</td>
<td>0.09</td>
<td>34.35</td>
</tr>
<tr>
<td><a href="http://www.freemarkettfoundation.com">www.freemarkettfoundation.com</a></td>
<td>3245</td>
<td>0.10</td>
<td>235165</td>
<td>0.17</td>
<td>72.47</td>
</tr>
<tr>
<td><a href="http://www.edri.org.et">www.edri.org.et</a></td>
<td>2390</td>
<td>0.07</td>
<td>19466</td>
<td>0.01</td>
<td>8.14</td>
</tr>
<tr>
<td><a href="http://www.africafoicentre.org">www.africafoicentre.org</a></td>
<td>2377</td>
<td>0.07</td>
<td>74462</td>
<td>0.05</td>
<td>31.33</td>
</tr>
<tr>
<td><a href="http://www.cde.org.za">www.cde.org.za</a></td>
<td>2133</td>
<td>0.06</td>
<td>69643</td>
<td>0.05</td>
<td>32.65</td>
</tr>
<tr>
<td><a href="http://www.kippra">www.kippra</a></td>
<td>2068</td>
<td>0.06</td>
<td>2704</td>
<td>0.00</td>
<td>1.31</td>
</tr>
<tr>
<td><a href="http://www.eeaecon.org">www.eeaecon.org</a></td>
<td>1939</td>
<td>0.06</td>
<td>120585</td>
<td>0.09</td>
<td>62.19</td>
</tr>
<tr>
<td><a href="http://www.separc.co.sz">www.separc.co.sz</a></td>
<td>1616</td>
<td>0.05</td>
<td>21178</td>
<td>0.02</td>
<td>13.11</td>
</tr>
<tr>
<td><a href="http://www.drpcngr.org">www.drpcngr.org</a></td>
<td>1501</td>
<td>0.05</td>
<td>6294</td>
<td>0.00</td>
<td>4.19</td>
</tr>
<tr>
<td>wiser.wits.ac.za</td>
<td>1277</td>
<td>0.04</td>
<td>17496</td>
<td>0.01</td>
<td>13.70</td>
</tr>
</tbody>
</table>

**Inlinks to TTs in sub-Saharan Africa**

Inlink counts refer to the number of links to each site crawled from the other sites crawled in the same project. The inlinks are often used to depict “sitations”, which is used as equivalent of citations in scholarly publishing. As Xing & Chu (2006: 3) explain, “content-based hyperlinks are commonly regarded as bibliographic citations or citations in short”. The term *sitation* was coined by Rousseau in 1997 (Rousseau 1997). Various authors (e.g. Bar-Ilan 2010; Prime, Bassecoulard & Zitt 2002), too, have likened hypertext links to citations. As a result, the TTs’ inlinks in this study reflected the web impact of the TTs. Table 2 provides the inlinks to the TTs’ web pages, directories, domains or sites from each other. The most inlinked TT in terms of page inlink counts was FMF, which received a total of 2890 page inlinks. The Foundation also received 1207 directory inlinks, 2 domain inlinks and 2 site inlinks.
inlinks. The rest of the TTs received less than 500 page inlinks and/or directory inlinks from the other TTs.

Table 2. Inlinks and outlinks per TT in sub-Saharan Africa

<table>
<thead>
<tr>
<th>Name</th>
<th>Page inlinks</th>
<th>Directory inlinks</th>
<th>Domain inlinks</th>
<th>Site inlinks</th>
</tr>
</thead>
<tbody>
<tr>
<td>freemarketfoundation.com</td>
<td>2890</td>
<td>1207</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ipss-addis.org</td>
<td>461</td>
<td>460</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>issafrica.org</td>
<td>267</td>
<td>266</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>fanrpan.org</td>
<td>236</td>
<td>178</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>atpsnet.org</td>
<td>114</td>
<td>109</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>saiia.org.za</td>
<td>50</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>uneca.org</td>
<td>30</td>
<td>17</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>idasa.org</td>
<td>22</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>aercafrica.org</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>.efdinitiative.org/ethiopia</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>acetafrica.org</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>suddinstitute.org</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>accord.org.za</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>esrftz.org</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>riftvalley.net</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>bidpa.bw</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ippr.org.na</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ai.org.za</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>heritageinstitute.org</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>kippra.org</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ccr.org.za</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>isser.edu.gh</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>irenkenya.com</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>minds-africa.org</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Outlinks of the sub-Saharan African TTs

Outlink counts refer to the number of links from each site crawled to other sites crawled within a given project; outlinks are the opposite of inlinks. Outlinks are like references or in-text citations in scholarly publications whereby the website offering outlinks is likened to a citing publication (Bjorneborn & Ingwersen 2001: 65). Table 3, which provides the number of outlinks for each TT that produced at least one (1) site outlink, shows that the IRR yielded the highest number of outlinks to the rest of the TTs covered in this study. The number of IRR’s page outlinks was the only one that surpassed two thousand outlinks while the rest of the TTs posted less than one thousand page outlinks each. The number of domain and site outlinks was highest in Food, Agriculture and Natural Resources Policy Analysis Network (FARNPNAN) of South Africa.
Table 3: Outlinks per TT in sub-Saharan Africa

<table>
<thead>
<tr>
<th>Name</th>
<th>Page outlinks</th>
<th>Directory outlinks</th>
<th>Domain outlinks</th>
<th>Site outlinks</th>
</tr>
</thead>
<tbody>
<tr>
<td>irr.org.za</td>
<td>2889</td>
<td>1206</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>.ipss-addis.org/new-ipss/</td>
<td>462</td>
<td>461</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>accord.org.za</td>
<td>260</td>
<td>260</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>.fanrpan.org/countries/angola/</td>
<td>185</td>
<td>175</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>.atpsnet.org/national_chapters/tanzania/</td>
<td>114</td>
<td>109</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>esrftz.org</td>
<td>99</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>fanrpan.org</td>
<td>33</td>
<td>21</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>isodec.org.gh</td>
<td>22</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>riftvalley.net</td>
<td>9</td>
<td>7</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>edri.org.et</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>aercafrica.org</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>saiia.org.za</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>uongozi.or.tz</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>uneca.org</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>mefmi.org</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ai.org.za</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>heritageinstitute.org</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>cres-sn.org</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>.efdinitiative.org/ethiopia</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>bidpa.bw</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>wiser.wits.ac.za</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4: Correlation between inlinks and outlinks of TTs in sub-Saharan Africa (Key: $r$ – Pearson correlation coefficient; sig – significance [2-tailed])

<table>
<thead>
<tr>
<th></th>
<th>Page inlinks</th>
<th>Directory inlinks</th>
<th>Domain inlinks</th>
<th>Site inlinks</th>
<th>Page outlinks</th>
<th>Directory outlinks</th>
<th>Domain outlinks</th>
<th>Site outlinks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page inlinks</td>
<td>$r$</td>
<td>1</td>
<td>0.970**</td>
<td>0.191</td>
<td>-0.059</td>
<td>-0.074</td>
<td>-0.091</td>
<td>-0.079</td>
</tr>
<tr>
<td>Sig</td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.280</td>
<td>0.405</td>
<td>0.739</td>
<td>0.679</td>
<td>0.609</td>
</tr>
<tr>
<td>Directory inlinks</td>
<td>$r$</td>
<td>1</td>
<td>0.330</td>
<td>0.188</td>
<td>-0.073</td>
<td>-0.091</td>
<td>-0.088</td>
<td>-0.070</td>
</tr>
<tr>
<td>Sig</td>
<td></td>
<td></td>
<td>0.056</td>
<td>0.286</td>
<td>0.680</td>
<td>0.609</td>
<td>0.620</td>
<td>0.693</td>
</tr>
<tr>
<td>Domain inlinks</td>
<td>$r$</td>
<td>1</td>
<td>0.847**</td>
<td>-0.187</td>
<td>-0.214</td>
<td>-0.007</td>
<td>0.077</td>
<td>0.666</td>
</tr>
<tr>
<td>Sig</td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.291</td>
<td>0.223</td>
<td>0.970</td>
<td>0.666</td>
<td></td>
</tr>
<tr>
<td>Site inlinks</td>
<td>$r$</td>
<td>1</td>
<td>-0.184</td>
<td>-0.209</td>
<td>0.078</td>
<td>0.174</td>
<td>0.325</td>
<td></td>
</tr>
<tr>
<td>Sig</td>
<td></td>
<td></td>
<td>0.297</td>
<td>0.236</td>
<td>0.660</td>
<td>0.325</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Page outlinks</td>
<td>$r$</td>
<td>1</td>
<td>0.969**</td>
<td>0.047</td>
<td>0.000</td>
<td>0.790</td>
<td>0.991</td>
<td>1</td>
</tr>
<tr>
<td>Sig</td>
<td></td>
<td></td>
<td>0.000</td>
<td>0.141</td>
<td>0.012</td>
<td>0.012</td>
<td>0.947</td>
<td></td>
</tr>
<tr>
<td>Directory outlinks</td>
<td>$r$</td>
<td>1</td>
<td>0.426</td>
<td>1</td>
<td>0.426</td>
<td>0.947</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig</td>
<td></td>
<td></td>
<td>0.426</td>
<td>1</td>
<td>0.947</td>
<td></td>
<td></td>
<td>0.000</td>
</tr>
<tr>
<td>Domain outlinks</td>
<td>$r$</td>
<td>1</td>
<td>0.911**</td>
<td>0.000</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Site outlinks</td>
<td>$r$</td>
<td>1</td>
<td>0</td>
<td>0.000</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).
In order to assess the relationship between the TTs’ in- and out-linking patterns, a correlational analysis of the number of inlinks and outlinks was performed using Pearson correlation analysis (embedded in the SPSS software) and yielded the results in table 4. Whereas there were high scores of correlation between the page and directory inlinks, there were mixed patterns of linking domains and sites. Strong and significantly correlated relationships occurred between page inlinks and directory inlinks ($r=0.970$), domain inlinks and site inlinks ($r=0.847$), page outlinks and directory outlinks ($r=0.969$) and domain outlinks and site outlinks ($r=0.911$), all at a significance level of $p \leq 0.01$. The majority of the relationships were either negatively or positively correlated but not significant at $p \leq 0.01$. The strongly correlated relationships were between two-like variables, that is, an inlink against another or an outlink against another. Relationships between inlinks and outlinks were either weak or very weak (i.e. $r$ was between $\pm 0.001$ and $\pm 0.299$ at $p \geq 0.01$).

Table 5 shows the number of file links originating from one site to another. The file interlinking patterns reveal that the IRR had 2889 outlinks to FMF while the outlinks from IPSS to IPSS were 461. The other top ranking file interlinking patterns were as follows: ACCORD to ISSAFRICA (260), FANRPAN to FANRPAN (185) and ATPSNET to ATPSNET (114). Three of the aforementioned file interlinkages were self-links, while only two were external links. The rest of the file interlinks were external links, that is, from one crawled site to another crawled site within the project. In considering the outlinking and inlinking countries (i.e. countries from which the links originated to the countries to which the links are directed), we observed that there were some cross-border interlinkages as shown in table 5.

In terms of country-to-country interlinkages, South Africa (ZAF) provided links to seven countries as follows, in the order of number of outlinking TTs; number of links: Angola [AGO] (1; 185), Ethiopia [ETH] (2; 21), Botswana [BWA] (1; 2), Kenya [KEN] (3; 4), Nigeria [NGA] (1; 2), Tanzania [TZA] (1; 2), and Ghana [GHA] (1; 1). Tanzania’s outlinks were directed to ZAF (3; 100), ETH (1; 2), and GHA (1; 1). Somalia (SOM) had two file links to KEN while Kenya’s fourteen file links targeted ETH (1; 5), SOM (1; 1), South Sudan [SSD] (1; 3) and ZAF (1; 5). Ghana provided twenty two file links from one TT to ZAF. On the part of Ethiopia, three TTs received a total of five links while BWA provided one file link to ZAF. It was also observed that some branches of some of the TTs provided links to the parent or other branches across the boarders (e.g. FANRPAN in Angola which provided 185 links to South Africa’s FARNPAN).

In-country (i.e. file links within the same country) file interlinkages occurred in the following countries in the order of number of TTs and number of links: ZAF (9; 3157), TZA (1; 114), and ETH (4; 467). Similar patterns were witnessed in the directory and domain interlinkages, although there were differences in the volume of links whereby directory and domain links were fewer than the file links. If the inlinks within the same country is a reflection of the recognition of each TT by another, then the patterns of interlinking among the TTs imply that some TTs within some countries have no recognition of one another as there were no links between them. South African TTs were the most interlinked.
<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>From</th>
<th>To</th>
<th>No of file links</th>
</tr>
</thead>
<tbody>
<tr>
<td>irr.org.za</td>
<td>freemarketfoundation.com</td>
<td>ZAF</td>
<td>ZAF</td>
<td>2889</td>
</tr>
<tr>
<td>.ipss-addis.org/new-ipss/</td>
<td>ipss-addis.org</td>
<td>ETH</td>
<td>ETH</td>
<td>461</td>
</tr>
<tr>
<td>accord.org.za</td>
<td>issafrica.org</td>
<td>ZAF</td>
<td>ZAF</td>
<td>260</td>
</tr>
<tr>
<td>.fanrpan.org/countries/angola/</td>
<td>fanrpan.org</td>
<td>ZAF</td>
<td>AGO</td>
<td>185</td>
</tr>
<tr>
<td>.atpsnet.org/national_chapters/tanzania/</td>
<td>atpsnet.org</td>
<td>TZA</td>
<td>TZA</td>
<td>114</td>
</tr>
<tr>
<td>esrftz.org</td>
<td>fanrpan.org</td>
<td>TZA</td>
<td>ZAF</td>
<td>50</td>
</tr>
<tr>
<td>esrftz.org</td>
<td>saia.org.za</td>
<td>TZA</td>
<td>ZAF</td>
<td>49</td>
</tr>
<tr>
<td>isodec.org.gh</td>
<td>idasa.org</td>
<td>GHA</td>
<td>ZAF</td>
<td>22</td>
</tr>
<tr>
<td>fanrpan.org</td>
<td>unaeca.org</td>
<td>ZAF</td>
<td>ETH</td>
<td>20</td>
</tr>
<tr>
<td>aercafrica.org</td>
<td>unaeca.org</td>
<td>KEN</td>
<td>ETH</td>
<td>5</td>
</tr>
<tr>
<td>riftvalley.net</td>
<td>issafrica.org</td>
<td>KEN</td>
<td>ZAF</td>
<td>5</td>
</tr>
<tr>
<td>edri.org.et</td>
<td>.efdinitiative.org/ethiopia</td>
<td>ETH</td>
<td>ETH</td>
<td>4</td>
</tr>
<tr>
<td>riftvalley.net</td>
<td>suddinstitute.org</td>
<td>KEN</td>
<td>SSD</td>
<td>3</td>
</tr>
<tr>
<td>mefmi.org</td>
<td>aercafrica.org</td>
<td>ZWE</td>
<td>KEN</td>
<td>3</td>
</tr>
<tr>
<td>fanrpan.org</td>
<td>bidpa.bw</td>
<td>ZAF</td>
<td>BWA</td>
<td>2</td>
</tr>
<tr>
<td>fanrpan.org</td>
<td>aercafrica.org</td>
<td>ZAF</td>
<td>KEN</td>
<td>2</td>
</tr>
<tr>
<td>fanrpan.org</td>
<td>ippr.org.za</td>
<td>ZAF</td>
<td>NGA</td>
<td>2</td>
</tr>
<tr>
<td>fanrpan.org</td>
<td>esrftz.org</td>
<td>ZAF</td>
<td>TZA</td>
<td>2</td>
</tr>
<tr>
<td>edri.org.et</td>
<td>acetforafrica.org</td>
<td>ETH</td>
<td>GHA</td>
<td>2</td>
</tr>
<tr>
<td>uongozi.or.tz</td>
<td>unaeca.org</td>
<td>TZA</td>
<td>ETH</td>
<td>2</td>
</tr>
<tr>
<td>unaeca.org</td>
<td>acetforafrica.org</td>
<td>ETH</td>
<td>GHA</td>
<td>2</td>
</tr>
<tr>
<td>cres-sn.org</td>
<td>aercafrica.org</td>
<td>SEN</td>
<td>KEN</td>
<td>2</td>
</tr>
<tr>
<td>heritageinstitute.org</td>
<td>riftvalley.net</td>
<td>SOM</td>
<td>KEN</td>
<td>2</td>
</tr>
<tr>
<td>bidpa.bw</td>
<td>fanrpan.org</td>
<td>BWA</td>
<td>ZAF</td>
<td>1</td>
</tr>
<tr>
<td>saiia.org.za</td>
<td>accord.org.za</td>
<td>ZAF</td>
<td>ZAF</td>
<td>1</td>
</tr>
<tr>
<td>saiia.org.za</td>
<td>ccr.org.za</td>
<td>ZAF</td>
<td>ZAF</td>
<td>1</td>
</tr>
<tr>
<td>saiia.org.za</td>
<td>issafrica.org</td>
<td>ZAF</td>
<td>ZAF</td>
<td>1</td>
</tr>
<tr>
<td>saiia.org.za</td>
<td>uneca.org</td>
<td>ZAF</td>
<td>ETH</td>
<td>1</td>
</tr>
<tr>
<td>ai.org.za</td>
<td>accord.org.za</td>
<td>ZAF</td>
<td>ZAF</td>
<td>1</td>
</tr>
<tr>
<td>ai.org.za</td>
<td>saiia.org.za</td>
<td>ZAF</td>
<td>ZAF</td>
<td>1</td>
</tr>
<tr>
<td>fanrpan.org</td>
<td>kippra.org</td>
<td>ZAF</td>
<td>KEN</td>
<td>1</td>
</tr>
<tr>
<td>fanrpan.org</td>
<td>ai.org.za</td>
<td>ZAF</td>
<td>ZAF</td>
<td>1</td>
</tr>
<tr>
<td>fanrpan.org</td>
<td>freemarketfoundation.com</td>
<td>ZAF</td>
<td>ZAF</td>
<td>1</td>
</tr>
<tr>
<td>fanrpan.org</td>
<td>isser.edu.gh</td>
<td>ZAF</td>
<td>GHA</td>
<td>1</td>
</tr>
<tr>
<td>fanrpan.org</td>
<td>irenkenya.com</td>
<td>ZAF</td>
<td>KEN</td>
<td>1</td>
</tr>
<tr>
<td>riftvalley.net</td>
<td>heritageinstitute.org</td>
<td>KEN</td>
<td>SOM</td>
<td>1</td>
</tr>
<tr>
<td>.ipss-addis.org/new-ipss/</td>
<td>unaeca.org</td>
<td>ETH</td>
<td>ETH</td>
<td>1</td>
</tr>
<tr>
<td>uongozi.or.tz</td>
<td>aercafrica.org</td>
<td>TZA</td>
<td>GHA</td>
<td>1</td>
</tr>
<tr>
<td>uongozi.or.tz</td>
<td>minds-africa.org</td>
<td>TZA</td>
<td>ZAF</td>
<td>1</td>
</tr>
<tr>
<td>uneca.org</td>
<td>aercafrica.org</td>
<td>ETH</td>
<td>GHA</td>
<td>1</td>
</tr>
<tr>
<td>.efdinitiative.org/ethiopia</td>
<td>uneca.org</td>
<td>ETH</td>
<td>ETH</td>
<td>1</td>
</tr>
<tr>
<td>wiser.wits.ac.za</td>
<td>issafrica.org</td>
<td>ZAF</td>
<td>ZAF</td>
<td>1</td>
</tr>
</tbody>
</table>
Networking of TTs in sub-Saharan Africa

Figure 1 shows a single network map, combining files on the websites of the TTs investigated in this study. The network map was developed using the Pajek software. The key players in the huge clusters marked A to H are the FMF (clusters A, B, and C). The files or pages that recorded the most number of outlinks belonged to the FMF. The files provided the most number of outlinks as follows: freemarketfoundation.com/view-event?1=107 (352); freemarketfoundation.com/view-event?1=103 (350); freemarketfoundation.com/view-event?1=104 (350); freemarketfoundation.com/view-event?1=105 (349); freemarketfoundation.com/view-event?1=111 (322); freemarketfoundation.com/view-event?1=100 (350); freemarketfoundation.com/view-event?1=95 (335). The rest of the pages provided less than 100 file links each.

Figure 1. Network map for TTs in sub-Saharan Africa

Most targeted sites by sub-Saharan African TTs

The external links, as explained in the methodology section, consists of a list of URLs targeting pages outside of each site. The links are rich in information which can point out the type of information that exist outside their own websites and which TTs deem to be important and valuable for their users. The URLs can partially provide the link analysers with information on the factors motivating web developers to provide links to external websites. Table 6 provides the list of the sites with the most number of links from the TTs’ websites. These links are targeting website developers (e.g. juizi.com), social networking sites (Facebook, LinkedIn, Twitter and YouTube), information access tools (e.g. online public access catalogue – OPAC), information databases (e.g. SABINET – South Africa Bibliographic and Information Network), and electronic e-mailing systems (e.g. mail.google.com), and more. All in all, a total of 5 047 772 external links were targeted by the sub-Saharan African TTs.

The network map, which was generated using the list of unselected external links, provides a partial revelation of the words that occurred twice or more in the hypertext links. There were 303 items that met the threshold. The words generated 121 clusters with 454 links and a “total link strength” of 551. A closer look at the top-level domains that were the most targeted shows that the .org (organization) was the most visible target with 418 occurrences, followed by .com (76) and .net (57). The presence of html, pdf and doc reveals the type of documents.
that have received links from the TTs while the regional names such as South Africa, Libya, Tanzania and Africa provide a glimpse of the type of geographic context of the type of information that is targeted by the TTs. The words publications, report, book, paper and articles as well as news and stories are some of the channels of information that are deemed important and therefore constitute the sources of information as well as channels of information dissemination for TTs and their targeted audiences. Examples of words that were indicative of the topics that were most regarded by the TTs, in the order of their frequencies of occurrence, include: business, women, agriculture, politics, law, youth, climate, economy, climate change, development, growth, jobs, peace, budget, democracy, partnership, security, and statistics, just to name those terms that occurred five or more times.

Table 6: Sites most targeted by TTs in sub-Saharan Africa

<table>
<thead>
<tr>
<th>No</th>
<th>Website/URL</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>juizi.com/</td>
<td>707841</td>
</tr>
<tr>
<td>2</td>
<td>.addthis.com/bookmark.php?v=250</td>
<td>590392</td>
</tr>
<tr>
<td>3</td>
<td>gtranslate.net/</td>
<td>514192</td>
</tr>
<tr>
<td>4</td>
<td>opac.eprc.or.ug/</td>
<td>514189</td>
</tr>
<tr>
<td>5</td>
<td>mail.google.com/</td>
<td>514187</td>
</tr>
<tr>
<td>6</td>
<td>.nitsak.com/</td>
<td>493411</td>
</tr>
<tr>
<td>7</td>
<td>.linkedin.com/company/the-afro-middle-east-centre</td>
<td>493410</td>
</tr>
<tr>
<td>8</td>
<td>.facebook.com/</td>
<td>493360</td>
</tr>
<tr>
<td>9</td>
<td>resolveuid/1775c217b63f4ae9b066f339d30dd7</td>
<td>53048</td>
</tr>
</tbody>
</table>

Figure 2: Most common words in ‘unselected’ external links

Conclusions and recommendations

The paper sought to assess the performance of TTs in sub-Saharan Africa on the Web. It was encouraging to note that the number of TTs that own websites in sub-Saharan Africa has
increased from 50% in 2010 (Mouton 2010) to 72.5% in 2016. Sadly, however, a sizable number of TTs, accounting for approximately 30%, still do not own websites. The institutions are likely to lose out on the benefits associated with the Web. Given the immense benefits associated with institutional presence and participation on the Web and other Internet-based applications such as social media technologies, and whereas the ranking of institutions has witnessed a paradigm shift whereby web presence, reach and impact are key indicators in the ranking systems, it is strongly recommended that the TTs which have not developed their own websites do so. The institutions’ Web presence may enhance their accountability status, both to the sponsors (e.g. government) and to the public.

Although the Think Tanks and Civil Societies Program (TTCSP) at the University of Pennsylvania highly ranks some of the TTs in sub-Saharan Africa, for example CODESRIA, the current study found that some of these TTs are not performing well on the Web in terms of the number of pages and even links. This pattern can be attributed to the language in which some of the websites are constructed and/or poor design of the affected TTs’ websites. Thirteen of the 91 TTs’ websites are constructed in a language other than English, a situation that might be limiting their Web visibility, reach and impact. For instance, it was noted that the most highly ranked TT whose website was constructed in non-English language was in position 28; all the top 27 positions were occupied by TTs whose websites are constructed in English language. Furthermore, the Webcrawler used to extract data from the websites did not recognize some pages and URLs, hence the poor performance of some of the TTs. Poorly designed and constructed websites may cause disasters to many organizations (Ahmad, Li & Azam 2005). The disasters may be in terms of loss in potential sales (Lu, Deng and Wang 2007). The TTs may be disadvantaged when it comes to the global ranking of institutions. It is therefore imperative that only professional web designers should be engaged to design and develop websites for TTs.

The analysis of the inlinks and outlinks demonstrates a limited activity among most TTs in sub-Saharan Africa. Only 24 and 21, out of the 66 TTs whose sides were crawled, received or provided links from and to one another, respectively. The number accounted for less than 40% in each case (i.e. inlinks and outlinks). This pattern may be taken to imply that there is limited knowledge sharing among the TTs. However, given that the web links alone cannot be an absolute indicator of the extent of knowledge sharing among institutions on the Web, there is a need for a study to investigate collaboration (including research collaboration) among the TTs in sub-Saharan Africa. Such a study will unearth the patterns, type and nature of collaborations among the TTs, a situation that may inform decisions on the management of TTs in the region. A content analysis of the internal links (links between pages in a site) nevertheless revealed that the most common usage of the sites lies in the organization of events (e.g. freemarketfoundation.com/view-event), publication and dissemination of documents (including reports), and showcasing the services offered by the TTs.

The file links in table 5 demonstrates a preference for in-country linkages (links among TTs in the same country) while links to external sites (links to sites outside Africa) was equally prominent (see table 6). There are minimal inter-country linkages in sub-Saharan Africa. This pattern follows the patterns of research collaboration whereby “internal” (within same institution or country) and foreign collaborations (between African and foreign countries) have been reported (see Onyancha & Ocholla 2007; Onyancha & Maluleka 2011; Narvaez-Berthelemot, Russell, Arvanitis, Waast, & Gaillard 2002; Sooryamoorthy 2009). Similar findings, whereby African-based institutions provide or receive links from or to institutions within the same country as well as provide links to institutions outside Africa,
have been reported in Onyancha (2007), Onyancha and Ocholla (2008) and Sooryamoorthy (2009).

The external links largely belong to institutions (mainly companies, organizations or networks) that offer essential services (e.g. web design, e-mail services, or social networking) or information (e.g. news). The TTs in Africa have paid little attention to their counterparts in the world as exhibited to the absence of even one TT based in a foreign country in table 6. This scenario may be attributed to lack or minimal contact or collaboration between the African-based TTs and their counterparts in the rest of the world. Finally, the content analysis of the hypertext links reveals the main areas of focus by the TTs (see Figure 2). It will be interesting to find out if these areas correlate with the research areas focused on by the same institutions in the region. A study of the subject content of the research publications is therefore recommended to unearth the relationship between the content of the hypertext links and research.

References


### Appendix A: List of Think Tanks in Sub-Saharan Africa (Source: McGann 2015)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name of TT</th>
<th>CODE</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Kenya Institute for Public Policy Research and Analysis</td>
<td>KIPPRA</td>
<td>Kenya</td>
</tr>
<tr>
<td>2</td>
<td>IMANI Center for Policy and Education</td>
<td>ICPE</td>
<td>Ghana</td>
</tr>
<tr>
<td>3</td>
<td>Council for the Development of Social Science Research in Africa</td>
<td>CODESRIA</td>
<td>Senegal</td>
</tr>
<tr>
<td>4</td>
<td>Botswana Institute for Development Policy Analysis</td>
<td>BIDPA</td>
<td>Botswana</td>
</tr>
<tr>
<td>5</td>
<td>African Centre for the Constructive Resolution of Disputes</td>
<td>ACCORD</td>
<td>South Africa</td>
</tr>
<tr>
<td>6</td>
<td>South African Institute of International Affairs</td>
<td>SAIJA</td>
<td>South Africa</td>
</tr>
<tr>
<td>7</td>
<td>Africa Institute of South Africa</td>
<td>AISA</td>
<td>South Africa</td>
</tr>
<tr>
<td>8</td>
<td>Centre for Conflict Resolution</td>
<td>CCR</td>
<td>South Africa</td>
</tr>
<tr>
<td>9</td>
<td>Centre for Policy Analysis</td>
<td>CEPA</td>
<td>Ghana</td>
</tr>
<tr>
<td>10</td>
<td>Institute for Security Studies</td>
<td>ISS</td>
<td>South Africa</td>
</tr>
<tr>
<td>11</td>
<td>African Economic Research Consortium</td>
<td>AERC</td>
<td>Kenya</td>
</tr>
<tr>
<td>12</td>
<td>Ghana Center for Democratic Development</td>
<td>CDD</td>
<td>Ghana</td>
</tr>
<tr>
<td>13</td>
<td>Food, Agriculture and Natural Resources Policy Analysis Network</td>
<td>FANRPAN</td>
<td>South Africa</td>
</tr>
<tr>
<td>14</td>
<td>Centre for Development and Enterprise</td>
<td>CDE</td>
<td>South Africa</td>
</tr>
<tr>
<td>15</td>
<td>Ethiopian Development Research Institute</td>
<td>EDRI</td>
<td>Ethiopia</td>
</tr>
<tr>
<td>16</td>
<td>Economic Policy Research Center</td>
<td>EPRC</td>
<td>Uganda</td>
</tr>
<tr>
<td>17</td>
<td>Ethiopian Economics Association</td>
<td>EEA</td>
<td>Ethiopia</td>
</tr>
<tr>
<td>18</td>
<td>REPOA, FKA Research on Poverty Alleviation</td>
<td>REPOA</td>
<td>Tanzania</td>
</tr>
<tr>
<td>19</td>
<td>Institute of Economic Affairs</td>
<td>IEA</td>
<td>Ghana</td>
</tr>
<tr>
<td>20</td>
<td>Institute for Global Dialogue</td>
<td>IGD</td>
<td>South Africa</td>
</tr>
<tr>
<td>21</td>
<td>Free Market Foundation</td>
<td>FMF</td>
<td>South Africa</td>
</tr>
<tr>
<td>22</td>
<td>Nigerian Institute of International Affairs</td>
<td>NIIA</td>
<td>Nigeria</td>
</tr>
<tr>
<td>23</td>
<td>Advocates Coalition for Development and Environment</td>
<td>ACODE</td>
<td>Uganda</td>
</tr>
<tr>
<td>24</td>
<td>African Center for Economic Transformation</td>
<td>ACET</td>
<td>Ghana</td>
</tr>
<tr>
<td>25</td>
<td>Institute of Economic Affairs</td>
<td>IEA</td>
<td>Kenya</td>
</tr>
<tr>
<td>No.</td>
<td>Organization Name</td>
<td>Abbreviation</td>
<td>Country</td>
</tr>
<tr>
<td>-----</td>
<td>---------------------------------------------------------------------</td>
<td>--------------</td>
<td>---------------</td>
</tr>
<tr>
<td>26</td>
<td>South African Institute of Race Relations</td>
<td>IRR</td>
<td>South Africa</td>
</tr>
<tr>
<td>27</td>
<td>Organization for Social Science Research in Eastern and Southern Africa</td>
<td>OSSREA</td>
<td>Ethiopia</td>
</tr>
<tr>
<td>28</td>
<td>Centre Ivoirien de Recherches Economiques et Sociales</td>
<td>CRES</td>
<td>Cote d'Ivoire</td>
</tr>
<tr>
<td>29</td>
<td>Centre for Population and Environmental Development</td>
<td>CPED</td>
<td>Nigeria</td>
</tr>
<tr>
<td>30</td>
<td>Centre for Research and Technology Development</td>
<td>RESTECH</td>
<td>Kenya</td>
</tr>
<tr>
<td>31</td>
<td>African Technology Policy Studies Network</td>
<td>ATPPSNET</td>
<td>Kenya</td>
</tr>
<tr>
<td>32</td>
<td>Initiative for Public Policy Analysis</td>
<td>IPPA</td>
<td>Nigeria</td>
</tr>
<tr>
<td>33</td>
<td>Centre for Development Studies</td>
<td>CDS</td>
<td>Ghana</td>
</tr>
<tr>
<td>34</td>
<td>Institute of Statistical, Social and Economic Research</td>
<td>ISSER</td>
<td>Ghana</td>
</tr>
<tr>
<td>35</td>
<td>Rift Valley Institute</td>
<td>RVI</td>
<td>Kenya</td>
</tr>
<tr>
<td>36</td>
<td>Centre for the Study of the Economies of Africa</td>
<td>CSEA</td>
<td>Nigeria</td>
</tr>
<tr>
<td>37</td>
<td>Makerere Institute of Social Research</td>
<td>MISR</td>
<td>Uganda</td>
</tr>
<tr>
<td>38</td>
<td>Institute for Public Policy Research</td>
<td>IPPR</td>
<td>Namibia</td>
</tr>
<tr>
<td>39</td>
<td>Institute for Empirical Research in Political Economy</td>
<td>IERPE</td>
<td>Benin</td>
</tr>
<tr>
<td>40</td>
<td>Centre d’Etudes, de Documentation et de Recherche Economiques et Sociales</td>
<td>CEDRES</td>
<td>Burkina Faso</td>
</tr>
<tr>
<td>41</td>
<td>Justice and Human Rights Institute</td>
<td>JHRI</td>
<td>Ghana</td>
</tr>
<tr>
<td>42</td>
<td>Economic and Social Research Foundation</td>
<td>ESRF</td>
<td>Tanzania</td>
</tr>
<tr>
<td>43</td>
<td>Groupe de Recherche en Economie Appliquee et Theorique</td>
<td>GREAT</td>
<td>Mali</td>
</tr>
<tr>
<td>44</td>
<td>Inter-Region Economic Network</td>
<td>IREN</td>
<td>Kenya</td>
</tr>
<tr>
<td>45</td>
<td>Strategic Transformation and Policy Centre</td>
<td>STPC</td>
<td>Cape Verde</td>
</tr>
<tr>
<td>46</td>
<td>Centre d’Etudes de Politiques pour le Developpement</td>
<td>CEPOD</td>
<td>Senegal</td>
</tr>
<tr>
<td>47</td>
<td>Institute for Public Policy Analysis and Management</td>
<td>IPPAM</td>
<td>Nigeria</td>
</tr>
<tr>
<td>48</td>
<td>Programme de Troisieme Cycle Inter-universitaire en Economie</td>
<td>PTCI</td>
<td>Burkina Faso</td>
</tr>
<tr>
<td>49</td>
<td>Integrated Social Development Center</td>
<td>ISODEC</td>
<td>Ghana</td>
</tr>
<tr>
<td>50</td>
<td>Institute for Peace and Security Studies</td>
<td>IPSS</td>
<td>Ethiopia</td>
</tr>
<tr>
<td>51</td>
<td>African Technology Policy Studies Network</td>
<td>ATPS</td>
<td>Tanzania</td>
</tr>
<tr>
<td>52</td>
<td>Mapungubwe Institute for Strategic Reflection</td>
<td>MISTRA</td>
<td>South Africa</td>
</tr>
<tr>
<td>53</td>
<td>Development Research and Projects Centre</td>
<td>dRPC</td>
<td>Nigeria</td>
</tr>
<tr>
<td>54</td>
<td>Institute of Policy Analysis and Research</td>
<td>IPAR</td>
<td>Kenya</td>
</tr>
<tr>
<td>55</td>
<td>Centre de Recherches, d’Etudes et d’Appui a l’Analyse Economique a Madagascar</td>
<td>CREAM</td>
<td>Madagascar</td>
</tr>
<tr>
<td>56</td>
<td>Centre Autonome d’Etudes et de Renforcement des Capacites pour le Developpement au Togo</td>
<td>CADERDT</td>
<td>Togo</td>
</tr>
<tr>
<td>57</td>
<td>Nigerian Institute for Social and Economic Research</td>
<td>NISER</td>
<td>Nigeria</td>
</tr>
<tr>
<td>58</td>
<td>Macroeconomic and Financial Management Institute of Eastern and Southern Africa</td>
<td>MEFMI</td>
<td>Zimbabwe</td>
</tr>
<tr>
<td>59</td>
<td>Mandela Institute for Development Studies</td>
<td>MINDS</td>
<td>South Africa</td>
</tr>
<tr>
<td>60</td>
<td>UONGOZI Institute</td>
<td>UONGOZI</td>
<td>Tanzania</td>
</tr>
<tr>
<td>61</td>
<td>Food, Agriculture and Natural Resources Policy Analysis Network</td>
<td>FanRPAN</td>
<td>Angola</td>
</tr>
<tr>
<td>62</td>
<td>Swaziland Economic Policy Analysis and Research Centre</td>
<td>SEPARC</td>
<td>Swaziland</td>
</tr>
<tr>
<td>63</td>
<td>Institute for Policy Analysis and Research</td>
<td>IPAR</td>
<td>Rwanda</td>
</tr>
<tr>
<td>64</td>
<td>African Institute for Applied Economics</td>
<td>AIAE</td>
<td>Nigeria</td>
</tr>
<tr>
<td>65</td>
<td>Development Policy Research Unit</td>
<td>DPRU</td>
<td>South Africa</td>
</tr>
<tr>
<td>66</td>
<td>Africa Freedom of Information Centre</td>
<td>AFIC</td>
<td>Uganda</td>
</tr>
<tr>
<td>67</td>
<td>United Nations Economic Commission for Africa</td>
<td>UNeca</td>
<td>Ethiopia</td>
</tr>
<tr>
<td>68</td>
<td>Afro-Middle East Centre</td>
<td>AMEC</td>
<td>South Africa</td>
</tr>
<tr>
<td>69</td>
<td>Brunthurst Foundation</td>
<td>BHF</td>
<td>South Africa</td>
</tr>
<tr>
<td>70</td>
<td>Cellule d'Analyse de Politiques Economiques du Cires</td>
<td>CAPEC</td>
<td>Ivory Coast</td>
</tr>
<tr>
<td>71</td>
<td>Centre d'Analyse de Politiques Economiques et Sociales</td>
<td>CAPES</td>
<td>Burkina-Faso</td>
</tr>
<tr>
<td>72</td>
<td>Centre d’Etudes et de Renforcement des Capacités d’Analyse et de Plaidoyer</td>
<td>CERCAP</td>
<td>Mali</td>
</tr>
<tr>
<td>73</td>
<td>Centre for Democracy and Development</td>
<td>CDD</td>
<td>Nigeria</td>
</tr>
<tr>
<td>74</td>
<td>Consortium for Social and Economic Research</td>
<td>CSER</td>
<td>Kenya</td>
</tr>
<tr>
<td>75</td>
<td>Centre d'Etudes et de Recherche sur les Analyses et Politiques Economiques</td>
<td>CSER</td>
<td>Congo</td>
</tr>
<tr>
<td>76</td>
<td>Consortium pour la Recherche Economique en Afrique</td>
<td>CREA</td>
<td>Senegal</td>
</tr>
<tr>
<td>77</td>
<td>Environmental Economics Policy Forum for Ethiopia</td>
<td>EEPFE</td>
<td>Ethiopia</td>
</tr>
<tr>
<td>78</td>
<td>Heritage Institute for Policy Studies</td>
<td>HIIPS</td>
<td>Somalia</td>
</tr>
<tr>
<td>79</td>
<td>Institute for Democracy in South Africa</td>
<td>IDASA</td>
<td>South Africa</td>
</tr>
<tr>
<td>80</td>
<td>Center for Conflict Resolution</td>
<td>CCR</td>
<td>South Africa</td>
</tr>
<tr>
<td>81</td>
<td>Institute of Policy Analysis and Research</td>
<td>IPAR</td>
<td>Rwanda</td>
</tr>
<tr>
<td>82</td>
<td>Institute of Security Studies</td>
<td>IPSS</td>
<td>Ethiopia</td>
</tr>
<tr>
<td>83</td>
<td>Sudd Institute</td>
<td>SI</td>
<td>South Sudan</td>
</tr>
<tr>
<td>No.</td>
<td>Organization Name</td>
<td>Abbreviation</td>
<td>Country</td>
</tr>
<tr>
<td>-----</td>
<td>--------------------------------------------------------</td>
<td>--------------</td>
<td>---------------</td>
</tr>
<tr>
<td>84</td>
<td>Centre for Economic Transformation</td>
<td>CET</td>
<td>Ghana</td>
</tr>
<tr>
<td>85</td>
<td>Centre for the Study of Governance Innovation</td>
<td>CSGI</td>
<td>South Africa</td>
</tr>
<tr>
<td>86</td>
<td>Tanzania Natural Resources Forum</td>
<td>TNRF</td>
<td>Tanzania</td>
</tr>
<tr>
<td>87</td>
<td>Center for Environment and Development</td>
<td>CED</td>
<td>Cameroon</td>
</tr>
<tr>
<td>88</td>
<td>Center for Policy Analysis</td>
<td>CEPA</td>
<td>Ghana</td>
</tr>
<tr>
<td>89</td>
<td>Centro Terra Viva</td>
<td>CTV</td>
<td>Mozambique</td>
</tr>
<tr>
<td>90</td>
<td>CODESMA</td>
<td>CODESMA</td>
<td>Senegal</td>
</tr>
<tr>
<td>91</td>
<td>Wits Institute for Social and Economic Research</td>
<td>WISER</td>
<td>South Africa</td>
</tr>
</tbody>
</table>
Bibliometrics to Knowledgometrics: Theory Method and Principle of Metrics Science

S.L. Sangam
Visiting Professor DRTC Indian Statistical Institute, Bangalore
Formerly UGC Emeritus Fellow
Dean Faculty of Social Science
Professor and Chairman
Department of Library and Information Science
Karnatak University, Dharwad – 580 003
E-mail: slsangam@gmail.com

Abstract
The terms librametry, bibliometrics, scientometrics, informetrics, webometrics and knowledgometrics refer to fields related to the study of the dynamics of disciplines as reflected in the production of their literature. These terms are used to describe similar and overlapping methodologies. The origins, historical survey, scope, application and development of each of these terms are presented in this paper.

Keywords: Bibliometrics, Scientometrics, Knowledgometrics

Conference topic: The theory method and principle of five metrics science concepts.

Introduction
In recent days, statistics has been applied to a number of areas such as perspective planning, industrial and agricultural development, etc. Library and Information managers have adopted a number of quantitative methods in order to evaluate library resources and services more objectively and effectively. Scientometrics/ Informetrics refer to quantitative techniques applicable to measure the records of human communicated literature. Over the years, several new terms have appeared on the horizon representing quantitative studies in library and Information Science.

The History of comparative Anatomy Part-I: A Statistical Analysis by Cole and Eale (1917) is considered to be the first bibliometric study in 1917. Hulme (1923) was the first to use the expression ‘statistical bibliography’ in 1923; later several studies have been conducted. Gross and Gross (1927) study is considered to be the third study in the field based on citations. After Hulme, the term statistical bibliography was used by Henkle in 1938 in his article “The periodical literature of Biochemistry” and Gosnell (1944, 1943) in his dissertation in 1943, and later in his article 1944. The historical development of the term statistical bibliography has been traced by Witting (1978) in a foot note. As the term was considered very clumsy, not very descriptive, and can be confused with statistics itself or bibliographies on statistics.

Librametrics
In 1948 at the Aslib’s (1949) conference in Lamington Spa, Ranganathan introduced the term Librametry for the first time. He suggested to develop librametry on the lines of biometry, econometry, and psychometry. His suggestions were avidly welcomed at the conference by Bernal and others. The term Librametrics has two roots: Libra and
**Metry.** The word ‘Libra’ connotes ‘library’ and ‘metrics’ means measurement. Further, as the librarian of the Madras University Library, he practiced various librametric techniques way back in 1925, in order to solve day today library problems and to streamline the day-to-day library activities, services for their clientele and also for the betterment of library professional as a whole.

The scope of the library is limited to the quantitative study of books, readers and staff. Here the books, readers and staff are the three constituent elements or factors of the library. The absence of any one of the three will make the library cease to exist. Each has its own potentiality and it is only a sum of the three that makes a library. Thus we can measure all the Characteristics of books, readers and staff.

Ranganathan (1969) in his paper presented in the DRTC 7th Annual Seminar (1969) suggested a few examples of statistics to library science. Based on his experience and the experiments carried out at the Documentation Research and Training Centre outlined the applications for Librametrics:

1. Determination of the strength of library staff;
2. Disposition of library staff for circulation work during different library hours;
3. Disposition of library staff for reference service during different library hours
4. Organization of Library System; Establishing the distinction between “service library” and “dormitory library”;
5. Design of library building, fittings, and furniture;
6. Book selection;
7. Absolute syntax and facet syntax in relation to classification;
8. Length of class number; and
9. Variation in style in writing catalogue entries; and

The librametrics studies if developed properly as suggested by Ranganathan could become a good indicator for measuring various activities of librarianship both quantitatively and qualitatively. It is therefore imperative on the professional schools of library and information science to incorporate Librametrics as a foundation course. Such a step would help us to have an objective as systematic approach to the field of library and information science.

**Bibliometrics**

The term bibliometrics was first coined by Prichard (1989) in 1969 in preference to existing terminology ‘statistical bibliography’. The word “Bibliometrics” has two roots: ‘biblio’ and ‘metrics. The term ‘biblio’ is derived from the combination of Latin and Greek word ‘biblion’ equivalent to Bylos, meaning book, paper which in turn was derived from the word Bylos, a city of Phenonicia, a noted city for export trade in paper. The word ‘metrics’, on the other hand, indicates the science of meter, i.e., measurement and is derived either from Latin or Greek word ‘metricus’ or ‘metricos’ respectively, each managing measurement. This term was coined for the first time by Alan Pritchard. He used the term to describe all ‘studies which seek to quantify the process of written communication’. Fairthorne (1969) also defined it as 'the quantitative treatment of the properties of recorded discourse and behaviour pertaining to it.'
Bibliometric studies include studies of the growth of the literature in some subject, how much literature is contributed by various individuals, groups, or organisations or countries; how much exists in various languages; how the literature on some subject is scattered (e.g., over documentary types, language journals); and how quickly the literature on some subject becomes out-of-date (Studies of obsolescence). Another important group of bibliometric studies relates to what sources author cite. Day-by-day this study is attaining sophistication and complexity, having national, international and inter disciplinary character. The backbone of Bibliometrics lies in its sound theoretical foundation most effectively laid by some pioneers like Lotka (1926), Bradford (1934), Zipf (1949), Duck J de Sola Price (1963), Bookstein (1976), Mandelbrot (1952), Brookes (1968,1969a,1969b,19905), Garifield (1976a, 1976b), Egghe (1990) and many others, and their techniques are capable of throwing light on various complicated problems faced by information scientists to quantify the process of written communication. The bibliometric tools can be applied to

1. Studies related to scattering of articles
2. Geographical distribution, language-wise distribution, institution-wise distribution of articles
3. Age distribution of documents
4. Use of information storage and retrieval
5. Application, in the library use studies.
6. To study the trends in research and identifying the growth of literature.
7. To identify authorship trends in documents on various subjects.
8. To measure the utility of library services
9. To evaluate the library collection, etc.

These definitions of librametrics and bibliometrics show that librametrics primarily aims at the quantitative analysis of the management of libraries and bibliometrics is limited to recorded knowledge. The publication in both the fields suggests that in librametrics and bibliometrics, one examines the statistical distributions of the processes relating to the utilization of documents, Library staff, and Library users, to establish a theory for the structural aspects of library. Bibliometrics and librametrics may therefore be commonly defined as areas in which one studies ‘information processes and information handling in libraries and information centres by quantitatively analyzing the characteristics and behaviour of documents, library staff, and library users.’

Scientometrics

In the 1960s, particularly in Eastern Europe, the term “scientometrics” was used to denote “measurement of informatics process.” The term informatics was then widely used to mean “documentation / information handling activities;” obviously, there is not much difference between bibliometrics of the West and the scientometrics of the East Europe. The term Scientometrics originated as a Russian term for the application of quantitative methods to the history of science, which studies the quantitative aspects of science. It was suggested by Dolrov and Kormoni (1969), often used with same meaning as the bibliometrics to mean ‘the application of quantitative methods to history of science’. This term came into prominence with the founding of the journal
named ‘Scientometrics’ by T. Braun in 1977, originally published in Hungary and currently from Amsterdam, The Netherlands. Scientometrics used to mean communication process in science including socio-cultural aspects, and appears to be almost synonymous with science of science with more stress on its quantitative aspects. It is also used as a generic term for a system of knowledge, which endeavours to study the scientific (and technological) system by using a variety of approaches within the area of science and technology studies. Scientometrics is concerned with the quantitative features and characteristics of science and scientific research. Emphasis is placed on investigations in which the development and mechanism of science are studied by statistical mathematical methods. Scientometrics is now considered as a part of the sociology of science and is applied to science policy making. Thus Scientometrics involves studies in:

1. Sociology of Science,
2. History of science,
3. Growth of literature
4. Behaviour of of scientists,
5. Science indicators.

Derek John de Solla Price (1963) was treated as the father of scientometrics. He was a physicist, a historian of science, an information scientist and worked as a teacher of applied mathematics at Raffles College (which was to become part of the University of Singapore in 1948). It was there that he formulated his theory on the exponential growth of science, an idea that occurred to him when he noticed the growth in the Philosophical Transactions of the Royal Society between 1665 and 1850. He had the complete set in his home while Raffles College had its library built. Further, Garfield's contribution (1976a, 1976b) to scientometrics is quite significant; his contributions are evolved through his Science Citation Index. Merton (1952) also had his view on scientometrics.

**Informetrics**

Information, in its most restricted technical sense, is a sequence of symbols that can be interpreted as a message. Information can be recorded as signs, or transmitted as signals. Information is any kind of event that affects the state of a dynamic system. Conceptually, information is the message being conveyed. The English word was apparently derived from the Latin stem (*information*-) of the nominative (*information*): this noun is in its turn derived from the verb "informare" (to inform) in the sense of "to give form to the mind", "to discipline", "instruct", "teach":. Metrics means measuring. Informetrics is the study of quantitative aspects of information. This includes the production, dissemination and use of all forms of information, regardless of its form or origin.

According to Brookes (1990) the word ‘Informetrics’ was first proposed by Otto Nacke of West Germany in 1979. FID constituted a committee with this name and Nacke was its first Chairman. Rajan (Rajan &Sen 1985), the next Chairman of the Committee, reformulated the objectives of informetrics as to (Almi&Ing 1997) provide reliable data for research and development, policy-making, planning; to evaluate institutions, projects, articles, products, and other academic activities, and (to
identify or to develop the techniques to trace the origins and development of concepts. Ravichandra Rao (1985, 1993, 1996) in a short communication on “Informetrics vis-à-vis Bibliometrics: Scope and its Development”, mentioned that it is a field wherein the flow of information and behavior of information are analyzed, measured and quantitative relationships are established. It is a scientific field wherein the developments of measurement of impact of information are assessed continuously. Bibliometrics may therefore be treated as synonymous to informetrics having a scope to analyze quantitative characteristics of information. An FID Committee constituted with broadly defined objectives in the provision of research and technical data subsequently gave this name.

Third International conference on Informetrics was held in Bangalore in 1991. 'Informetrics’ was used as a generic term to mean “The use and development of a variety of measures to study and analyse several properties of information in general and documents in particular the study of the quantitative aspects of information in any form, not just records or bibliographies. Informetrics is the study of quantitative aspects of information. This includes the production, dissemination and use of all forms of information, regardless of its form or origin. As such, informetrics encompasses the fields of which studies quantitative aspects of science. It is mostly concerned with development of models to explain and identify the various characteristics of the literature. It also discusses scientific productivity, collaborative research, etc.

**Webometrics**

The science of webometrics (also cybermetrics) tries to measure the World Wide Web to get knowledge about the number and types of hyperlinks, structure of the World Wide Web and usage patterns. According to Björneborn and Ingwersen (2004), the definition of webometrics is "the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches." The term *webometrics* was first coined by Almind and Ingwersen (1997). A second definition of webometrics has also been introduced, "the study of web-based content with primarily quantitative methods for social science research goals using techniques that are not specific to one field of study" (Thelwall, 2009), which emphasizes the development of applied methods for use in the wider social sciences. The purpose of this alternative definition was to help publicize appropriate methods outside of the information science discipline rather than to replace the original definition within information science.

Cybermetrics is one of the recently emerged fields in the line of metric studies. It has gained much popularity since the mid-1990 with the advent of Information Technology. As it is mainly concerned with the computer-science-based approaches, it has superseded all other metric studies in this Internet Era. Cybermetrics is proposed as a generic term for “The study of the quantitative aspects of the construction and use of information resources, structures and technologies on the whole Internet drawing on bibliometric and informetric approaches.” Cybermetrics thus encompasses statistical studies of discussion groups, mailing lists, and other computer – mediated communication on the internet, including the www. Besides
covering all computer-mediated communication by using internet applications, this
definition of cybermetrics also covers quantitative measures of the internet backbone
technology, topology and traffic. The breadth of coverage of cybermetrics implies
large overlaps with proliferating computer-science-based approaches in analyses of
web contents, link structures, and web usage and web technologies. The Webometrics
which studies are the quantitative aspects of the World Wide Web. The Cybermetrics
which is similar to webometrics, but broadens its scope which include namely the
electronic resources. Research of all network-based communications by using
informetrics or other quantitative measures is called webometrics.

There has been a revolutionising symbiosis between computer and communication
technologies in the west over the past ten years. The invention of World Wide Web
(www) a part of the ‘INTERNET’, which is the mother of networks, has practically
webbed the information globally under less than one roof. There has been a shift in
navigational approaches from syntactical to semantic (i.e., from sentences to words),
as an ever increasing number of research institutes, universities and business
organisations are currently providing information about themselves in the form their
articles, publications, reports, catalogues and other information resources on the
INTERNET in general and the www in particular. This is now becoming the
accepted method of disseminating and sharing information resources in hypermedia.
Information science research has also changed, with much research to find out, how
the new technologies are being used, particularly e-mail and the web. In addition to
user studies there have been attempts to extract new kinds of information from the
web. Being a global document network initially developed for scholarly use, it is now
inhabited by a diversity of users, and the web constitutes an obvious research area for
bibliometrics, scientometrics and informetrics.

**Altmetrics**

Altmetrics or alternative metrics for measuring research impact emerged as a subfield
of webometrics (Priem, 2011). It is largely depend on social web has expand the
scope of webometrics. Application designers should continue to build systems to
display altmetrics, develop methods to detect and repair gaming, and create metrics
for use and reuse of data ultimately. With the setting up of a metric to be used for
performance assessment; researchers will modify their behaviour to optimize their
performance. By promoting their work through publishing, giving talks at meetings,
doing interviews, things like that, it helps raise awareness of their work and
apparently increases its impact. The focus of Altmetrics is the attention received by
their work in social media namely Facebook, Twitter, blogs, forums, micro blogs,
social networks, products/services reviews, virtual world and others. No one can
claim that Altmetrics is perfect but it is one way forward in the quest for perfection
(Sangam, 2015).

**Knowledgometrics**

Knowledge metrology is an interdisciplinary integrated research topic. "Knowledge
Metrology is the scientific evaluation of Wuhan University, China Research centre in
bibliometrics, information metrology, scientific metrology, network metrology etc. It
is having its own theory, method and application (1-4-2017). It further studies the
whole human knowledge system as research object, and study the social ability of knowledge, and the social relationship of knowledge through the quantitative analysis and computing technology. In addition to the important research of Wuhan University. Early in 1998, the discipline planning of the “knowledge metrology” was put forward in the International Symposium on Research Evaluation and Scientometrics held in Beijing, the idea oriented from the concept of “element of knowledge” created by Hongzhou Zhao (1999). Another notable study was The “element of knowledge” theory put forward by Hongzhou Zhao was published in Chinese Journal in 1990, and which has been re-visited in detail in Chapter 2 of Turning Points-The Nature of Creativity written by Chaomei Chen (2011) The emerging of “Knowledge domain mapping”, “Knowledge discovery” “text mining” used in knowledge management are all the important development in knowledgometrics. Apart from this, China has big contribution from WISE Lab and other Research Institutes and Universities.

**Theoretical Principles and Science Concepts**

Bibliometrics is concerned with theoretical and philosophical foundations. Some of the important studies on theoretical and philosophical foundations are in the area of:

- Law of Scattering (Bradford’s law)
- Author productivity (Lotka’s law)
- Word productivity (law of Least Efforts)
- Impact Factor of Journal
- H-index
- Success-breeds success phenomenon
- Circulation theory
- Information Product and Processes (IPP) and Duality in IPPs

There are other theoretical studies, especially in the area of circulation theory, citation analysis, sources-items relation, etc. Some of these are discussed in:

- Proceedings of the ISSI Conferences in Scientometrics and Informetrics (held bi-annually since 1987)

Studies based on citation data have several limitations. They are:

- Citation studies are mostly dependent on data from databases such as Web of Science, SCOPUS, etc. which cover only a limited number of journals and its
coverage does not remain constant since new journals are added regularly and some are dropped,

- Solutions to problems of eliminating self-citation are cumbersome.

A primary objective of bibliometric research is the development of a general and systematic set of theories from which hypotheses can be generated and tested. Scientometric studies vary from each other from several points of view. They adopt different methods of data collection as well as different techniques. Even there are no universally accepted terminologies. In addition, use of algebraic symbols varies from one study to another. Under these circumstances, it would be difficult to think of “bibliometric or scientometric standards” let alone formulating them.

Most of the scientometric studies are empirical in nature. In such circumstances, to reproduce the research, one has to repeat the survey and analyse the data right from the beginning. Even then, we may not get the same result. In natural sciences, it is possible and quite common that research may be repeated in laboratories. But in social sciences this is not only difficult, but is not possible. Further, an important cause of the overall unreliability and therefore a cause of invalidity in any basic research in the social sciences are due to space and time factors. It is therefore difficult to reproduce the results of research.

The fact, whether we call our research area as librarmetry or bibliometrics or scientometrics or informetrics most of the topics we deal with are:

- Quantitative aspects of library and information studies, especially use and user studies, growth of collection, age distribution of documents, circulation statistics, etc.
- Journal productivity (by coverage, by use, by citation, cost-effectiveness measures, impact factor, h-index, sources of citations, immediacy of citations, age of sources cited, coverage in databases, etc.)
- Measures of productivity or author productivity, including studies related to multiple authorship (number of publications, cost-effectiveness measures, impact factor, h-index, reprints request, photo copies made, sources of citations, immediacy of citations, number of reviews, adoption rates (text books) etc.)
- Obsolescence and growth of literature
- Co-citation, bibliographic coupling, co-word analysis, rank distribution of words, etc.
- Quantitative analyses of science (science indicators -- country-wise, language-wise, subject-wise, etc.)
- Identifying relationships among various disciplines, structure of subjects, etc.
- Evaluation of scientific research (by institutions, by individuals, by countries, etc.)

Acceptance of a single term to define a subject and acceptance of its scope are necessary for any scientist. Otherwise, it is difficult to include it in a syllabus. It is also difficult to get research grant from different agencies. It helps us in identifying
the research groups especially at national and international levels. At present, the term 'scientometrics' is used as synonym to both 'bibliometrics' and 'informetrics'.

In order to encourage communication and exchange of professional information in the field of scientometrics and informetrics, a Society, called 'the International Society for Informetrics and Scientometrics' (ISSI) was founded in 1993. It is an association of professionals active in the emerging interdisciplinary fields of informetrics, bibliometrics / scientometrics, technometrics and webometrics. Among its membership are scientists from over 30 countries representing all five continents. The Articles of Association state that the aim of ISSI is the advancement of the theory, methods and explanations through the following two main streams:

1. Quantitative Studies related to
   - Scientific, technological and other scholarly substantive information.
   - Science of science and technology, social sciences, arts and humanities'.
   - Generation, diffusion and use of information.
   - Information systems, including libraries, archives and databases


Associations/ Institutions/Organizations

The Society was founded at the International Conference on Bibliometrics, Informetrics and Scientometrics held in Berlin, 11-15 September in 1993. This conference was the fourth of a series of prominent biennial conference that subsequently have been held under the auspices of the Society. The first three earlier conferences were held in Diepenbeek, Belgium (1987, Chairman: Dr. Leo Egghe), London, Ontario, Canada (1989, Chairman: Dr J M Tague) and in Bangalore, India (1991, Chairman: Dr I K Ravichandra Rao). The Society was incorporated with formal Articles of Association in 1994 in the Netherlands (Utrecht.) Dr. Hildrun Kretschmer was elected its first President.

With Berlin as its virtual center COLLNET was set up on January 1st, 2000, under the leadership of Hildrun Kretschmer, in her capacity as coordinator. The network is to comprise the prominent scientists, who work at present mostly in the field of quantitative science studies, coming from 15 countries of America, Asia, Australia and Europe. The intention is to work together in co-operation both on theoretical and applied aspects. COLLNET conducts every year International Conference in scientometrics.

There are four important journals in scientometrics - Journal of Scientometrics, Journal of Informetrics and COLLNET Journal of scientometrics and Information Management, Journal of Scientometrics Research (edited by S. Bhattacharya, CSIR NISTADS India). These journals are publishing original studies, short communications, and preliminary reports, and review papers, letters to the editor and book reviews on scientometrics. Due to its fully interdisciplinary character, the journal is indispensable to research workers and research administrators. It provides valuable assistance to librarians and documentalists in central scientific agencies, ministries, research institutes and laboratories.
On 27th March 2011 at Tumkur in India, “Institute of Scientometrics” was founded by Prof. S. L. Sangam along with his research students, with an objective to promote research in Scientometrics. It is a virtual and non-profit organization. Much of the research works in this area were carried out in National Institute of Science Communication and Information Resources (NISCAIR, a wing of CSIR) (formerly known as INSDOC), National Institute of Science Technology and Development Studies (NISTADS, a wing of CSIR) and Documentation Research and Training Centre of the Indian Statistical Institute. In recent times, a number of research works were published from Departments of Library and Information Science of many Universities in India. In 2009, UGC recognized Department of Library and Information Studied of the Karnataka University, Dharwad a Special Assistant Program (SAP) in scientometrics.

**Challenges of future research in the field of Scientometrics**

Scientometrics has strong connections with economics and sociology of science as well as science policy. It has many challenges for future research:

1) Delineation and mapping of scientific areas;
2) Measuring multi-disciplinarity;
3) Positioning indicators have challenged the traditional input-output measures, prioritizing volumes, market shares and productivity;
4) Capturing creativity and innovation;
5) Growth-diversity models; and
6) Data de-mining, knowledge, flow measurements and diversity issues.

**Conclusion**

The field of cybermetrics approaches the boundaries of bibliometrics, because some activities in cyberspace normally are not recorded, but communicated synchronously as in chat rooms. Cybermetric studies of such activities still fit in the generic field of Informetrics as the study of the quantitative aspects of information “in any form” and “in any social group” as stated by Tague-Sutcliffe (1992). The inclusion of webometrics expands the field of bibliometrics, as webometrics inevitably will contribute to further methodological developments of web-specific approaches. As ideas rooted in bibliometrics, scientometrics and Informetrics contributed to the emergence of webometrics, ideas in webometrics might also contribute to the development of these all-embracing fields.

**References**


Aslib proceedings. 1, 1949, 102.

Brookes, B.C. (1969a). The complete Bradford-Zipf bibliograph. Journal of Documentation. 24, 1; S8-
Gosnell, Chas F. The rate of obsolescence in college library book collections as determined by the analysis of the three select list of Books for College Libraries (Dissertation, Newyork University, 1943)
Priem, Jason; Taraborelli, Dario; Groth, Paul; Neylon, Cameron(2011). "Altmetrics: A manifesto (v 1.01)". Altmetrics.


Ravichandra Rao, I.K. (J981). Documents and user distribution in transaction records of Canadian university libraries. Ph. D Thesis. Faculty of Graduate Studies, School of Library and Information Science; the University of Western Ontario, London, Ontario, Canada.


Review of the Research on Knowledge Metrology - Bibliometrics (Retrived on 03-04-2017)

Wuhan University Library and Information Science Institute (Retrived on 03-04-2017)

Liu Zeyuan and Chinese Knowledge Metrology (Retrived on 03-04-2017)
Assessing Research and its Impacts: The *Generalized* Implementation Problem and a *Doubly-Conditional* Performance Evaluation Model

Cinzia Daraio

daraio@dis.uniroma1.it
Department of Computer, Control and Management Engineering “Antonio Ruberti” (DIAG), Sapienza University of Rome, Rome (Italy)

Abstract

This paper addresses the issue of designing relevant models of indicators to assess research and its impacts. The evaluation of research activities is a complex task for many reasons. There are no perfect indicators or metrics which fit for all purposes. In order to understand the appropriateness of the indicators to be used, we need to frame the problem taking into account the *systemic* nature of the phenomena and to develop *models* of metrics that are as close as possible to the reality being assessed. We show some examples of the usefulness of such a framework by discussing university rankings, the complexity of the assessment of research through the *generalized* implementation problem and presenting a *doubly conditional* performance evaluation model.

Conference Topic
Indicators, Science policy and research assessment, Science policy (on different levels)

Introduction: an overall framework for the assessment of Research and its Impacts

According to Daraio (2017a), each metric of research assessment (in this paper metrics and indicators are used as synonym) is based on a *model* that can be implicitly or explicitly defined and discussed. If the model underlying the metric is not described, this does not mean that the indicator is more robust to modelling choice. It simply means that you do not explicitly clarify and account for the underlying theoretical choices, methodological assumptions and data limits. Thus, as a consequence, if you do not specify your model of the metric, you may not check its robustness.

Developing models is important for two main reasons: i) to learn about the explicit consequences of assumptions, test the assumptions, highlight relevant relations; ii) to improve, to better operate, document/verify the assumptions, decompose analysis and synthesis, systematize the problem and the evaluation/choice done, explicit the dependence of the choice to the scenario. There are however several pitfalls and difficulties in modelling, which mainly relate to the possibility that the targets are not quantifiable; the complexity, uncertainty and changeability of the environment in which the controlled system works and, the limits in the decision context; the intrinsic complexity of calculation of the objective of the analysis.

Within this context, Daraio (2017a) proposes a *framework*, intended as a background which includes the main conditions, circumstances, ideas, and so on, to the realization of activities related to the assessment of research and its impacts. It is suggested as a reference to develop models of metrics accounting for the systemic nature of the research activity and its interrelations with teaching and innovation activities. See Figure 1 for an overall illustration.
Figure 1. A systemic framework for the development of models of metrics (source: Daraio 2017a)

Table 1 shows an outline of the definition of the different dimensions and components of the framework illustrated in Figure 1. The reader is referred to Daraio (2017a) for additional details and references. Daraio (2017a) states that “the ability to develop (and afterwards understand and use effectively) models for the assessment of research is linked and depends, among other factors, on the degree or depth of the conceptualization and formalization, in an unambiguous way, of the underlying idea of quality”.

The assessment of the research activity is also complicated by the quantification of data and its processing for their use in different contexts and for different purposes, including process monitoring, input-output monitoring, ex-ante and ex-post evaluation. In this case, according to Daraio and Glänzel (2016), there is a need to specify standards and rules for metadata definition and quantification (for additional details and references, see Daraio and Glänzel, 2016). Daraio (2017b) discusses the main challenges of econometric approaches to assess the productivity of research systems in this broader framework (characterized by Theory-Methodology and Data dimensions) and makes an attempt at standardization and harmonisation applied to the methodological dimension of the research assessment.

The main contribution of this paper is to show some examples of the usefulness of the framework introduced so far (Figure 1) for designing relevant models of indicators to assess research and its impacts. These examples include a discussion on

- university rankings,
- the complexity of the assessment of research through the generalized implementation problem,
- a doubly conditional performance evaluation model.

These issues are addressed in separate sections in the following. A final section summarizes and concludes the paper.
Table 1. Definition of the components of the framework

<table>
<thead>
<tr>
<th>Dimension component</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Theory</strong></td>
<td>Identifies the conceptual content of the analysis, answering the question “what is the domain of interest” and delineating the boundary of the investigation. Education and Research are the main conceptual blocks of Theory. Their interrelations and their complementarities should be considered in a systematic way when assessing research.</td>
</tr>
<tr>
<td><strong>2. Methodology</strong></td>
<td>Identifies the range of methods, techniques and approaches that are relevant for the evaluation of research. It answers the question “how” the investigation is handled. Efficiency, Effectiveness and Impact are the subject of the assessment. They go from the output (baseline) that is the result of the transformation of inputs in outputs; to efficiency which relates output to inputs with respect to an estimated efficient frontier, effectiveness considers inputs, output and account for the aims of the activity, while impact refers to all contributions of research outside Academia.</td>
</tr>
<tr>
<td><strong>3. Data</strong></td>
<td>Data are a relevant dimension often neglected in modelling building. Data have a problematic definition because this depends on their use not on inherent characteristics of the data (Borgman, 2015, p. 74). Data are instances coming from the domain of interest and represent the means by which the analysis are carried out. Availability, Interoperability and “Unit-free” property refer to the usability of data, alternatives and choices that affect the data. This is the way in which heterogeneous data systems are able to exchange information in a meaningful way.</td>
</tr>
<tr>
<td><strong>Quality</strong></td>
<td>“Fitness for purpose”. It is the overarching concept of the framework. It is also an attribute of the different dimensions of the framework.</td>
</tr>
<tr>
<td><strong>Implementation factors</strong></td>
<td>Tailorability: Adaptability to the features of the problem at hand. Transparency: Description of the choices made and underlying hypothesis masked in the proposed/selected theory/method/data combination.</td>
</tr>
<tr>
<td><strong>Enabling conditions</strong></td>
<td>Openness: Accessibility to the main elements of the modelling. Convergence: Evolution of the transdisciplinary approach, which allows for overcoming the traditional paradigms, increasing the dimensional space of thinking. Mixed Methods: Intelligent combination of qualitative and quantitative approaches. Knowledge infrastructures: Networks of people that interacts with artifacts, tools and data infrastructures.</td>
</tr>
</tbody>
</table>
An illustration on university rankings

In this section we briefly describe how the framework introduced in the previous section could be applied for the interpretation of university rankings.

Let us consider the interpretation of the university rankings debate in Nordic countries. Piro and Sivertsen (2016) try to explain the differences in the scores of universities on the basis of the different assumptions made, while, Elken, Hovdhaugen, and Stensaker (2016) found that "rankings have a relatively modest impact on decision-making and strategic actions in the Nordic universities studied, and that there are few signs of rankings challenging the existing identities of the universities in this region". The latter seems to support an alternative role of rankings in the national context (see also Douglass, 2016). On the other hand, Daraio, Bonaccorsi and Simar (2015)¹ found that Nordic universities, taking into account their multidimensional activities, perform in line with the European reference standard. What can the present framework tell us about this rankings comparison? We can frame the comparison of ranking results according to the three dimensions (theory, methodology and data) and their constituent building blocks. See Table 2.

Gingras (2016) identifies three essential characteristics of a good indicator namely: adequacy, sensitivity and homogeneity and afterwards discusses why some rankings and indicators are invalid measures and nonetheless continue to be used.² Daraio and Bonaccorsi (2017) suggest that it would be possible to go beyond university rankings through the intelligent integration of existing data that may lead to an open-linked data platform that may permit the construction of new indicators without designing the indicators on a custom basis.

However, Moed (2017) argues that “current [ranking] systems are still one-dimensional in the sense that they provide finalized, seemingly unrelated indicator values rather than offering a dataset and tools to observe patterns in multifaceted data.”

The generalized implementation problem in Research Assessment

The operationalization of the general framework of Figure 1 needs the discussion and extension (along the three dimensions of the framework) of the implementation problem which is typical of quantitative methods (mostly mathematical models) in Operations Research and Management Science (Schultz and Slevin, 1975; Hildebrandt, 1977; Bryant, 1989; Keys, 1991; Mingers and Gill, 1997).

¹ Daraio, Bonaccorsi and Simar (2015) describe the main criticisms of rankings that their proposed approach (based on conditional directional distance functions) addressed, namely: 1. Monodimensionality (universities perform several missions: teaching, research, third mission) 2. Statistical robustness (multiple rankings, probabilistic rankings, robustness of rankings) 3. Dependence on university size and subject mix (rankings favour old and large universities where scientific, technical and medical disciplines prevail) 4. Lack of consideration of the input-output structure (ignore the amount of resources used). They propose a conditional directional distance approach as a fair comparison: a novel two-stage approach of conditional efficiency scores (nonparametric location scale model) allowed them to estimate the managerial efficiency scores as the residuals (what remains after having eliminated size and disciplinary specialization effects). They propose a bootstrap based approach for estimating error bounds for managerial efficiency scores and hence provide some statistical robustness.

² In reviewing Gingras (2016)'s book, Schubert (2017) points out that “Adequacy can, of course, be interpreted only with respect to a given question or task. Actually, this is the only criterion an indicator should fulfil. Sensitivity, homogeneity or whatever else are technicalities which are of secondary importance as long as the indicator is adequate to its task.” See also the discussion reported in Zitt (2015).
Table 2. Example of Application on University Rankings

<table>
<thead>
<tr>
<th>Dimension component</th>
<th>Presence and degree of implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Theory</strong></td>
<td>NO: usually there is not a description of the overall conceptual notions underlying the proposed rankings</td>
</tr>
<tr>
<td>Education</td>
<td>NO</td>
</tr>
<tr>
<td>Research</td>
<td>YES</td>
</tr>
<tr>
<td>Innovation</td>
<td>NO</td>
</tr>
<tr>
<td><strong>2. Methodology</strong></td>
<td>YES</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Usually the outputs are mostly considered, while the efficiency, effectiveness and impact are not included</td>
</tr>
<tr>
<td>Impact</td>
<td></td>
</tr>
<tr>
<td><strong>3. Data</strong></td>
<td>NO: Generally the underlying data are not discussed at length</td>
</tr>
<tr>
<td>Availability</td>
<td>NO</td>
</tr>
<tr>
<td>Interoperability</td>
<td>NO</td>
</tr>
<tr>
<td>“Unit-free” property</td>
<td>NO</td>
</tr>
<tr>
<td>Quality</td>
<td>NO</td>
</tr>
<tr>
<td><strong>Implementation factors</strong></td>
<td>NOT AVAILABLE</td>
</tr>
<tr>
<td>Tailorability</td>
<td></td>
</tr>
<tr>
<td>Transparency</td>
<td>LOW (relates mainly to the methodology)</td>
</tr>
<tr>
<td>Openness</td>
<td>LOW (relates mainly to the methodology)</td>
</tr>
<tr>
<td><strong>Enabling conditions</strong></td>
<td></td>
</tr>
<tr>
<td>Convergence</td>
<td>NO</td>
</tr>
<tr>
<td>Mixed Methods</td>
<td>NO</td>
</tr>
<tr>
<td>Knowledge infrastructures</td>
<td>NO</td>
</tr>
</tbody>
</table>

It is based on the application of methods developed as basic research to concrete organizations and contexts for specific problem-solving. We refer to the generalized problem because we consider i) the interaction of method development with its useful application; ii) that the implementation changes the unit of assessment; iii) that knowledge and technological innovation may be at place. Moreover, needless to say, the identification of the right problem and the development of an appropriate model are crucial determinants of success.

Figure 2 shows the extension of the implementation problem along the main dimensions of the Framework reported in Figure 1.

Along the Methodology dimension, we observe an evolution of the implementation theory, thanks to the introduction of the multi-methodology approach (Mingers and Gill, 1997; Mingers, 2006). In Figure 3 (Panel A), on the top-left side there is an illustration of three systems, which constitute the context of the intervention. According to Mingers (2006) who extended the Checkland (1981)'s two systems (problem solving and problem content systems), the three systems are the intervention system (agents undertaking the intervention), the problem content system (real-world situation of concern) and the intellectual resources system (available theories and methodologies).

A fruitful connection is the work on cognitive biases and their impact on decision-making introduced by Tversky and Kahneman (1974) which present three heuristics that are employed in taking decisions under uncertainty:
- representativeness, related to the use of categories to evaluate the probability that a given event belongs to a given class or a given process;
- availability of instances or scenarios, related to the evaluation of the frequency of a given class or the likelihood of a particular progress; and
- adjustment from an anchor, related to numerical prediction, when a pertinent value is available. These heuristics that are generally effective, lead to regular and expected biases that should be taken into account.

Intervention systems are made by agents (managers, decision makers...) who build models of problem content system (including inputs, organization, goals, criteria etc.) with the purpose of problem solving and obtaining outputs (activities, projects, solutions) which generate effects that interact and influence the previous systems. The analysis of the problem and context of intervention can be very difficult to address globally. This is due to the consideration of the two additional dimensions of Data and Theory. Here the approach of the Levels of Abstraction (LoA) proposed by Floridi (2002, 2008) could be helpful for our implementation problem to move from Data to information. Floridi’s (2002, 2008) method of abstraction relies on a method borrowed from a branch of theoretical computer science called Formal Methods, in which discrete mathematics is used to specify and analyse the behavior of information systems.

The idea behind the method of Levels of Abstraction (Floridi 2008) is quite simple and straightforward: reality can be viewed from different levels. Perhaps the most crucial feature of the method of LoA is that the identification relation between two observables is never absolute. Rather, the identification is always contextual and the context is a function of the level of abstraction chosen for the required analysis (Floridi and Sanders 2004). A LoA can be defined as a finite non-empty set of observables (which are interpreted typed variables, i.e. typed variables together with a statement of what features of the system under consideration they stand for) which are expected to be the components of a theory (this in our set-up, is given by the theory-method-data paradigm). A gradient of abstraction is made by a collection of LoAs and represents a kind of interface. It is used to analyse some systems from varying points of view or at different LoA. Hence, the method permits the analysis of systems by means of models developed at specific gradients of abstraction (interfaces). Figure 3 (Panel A) reports an illustration of the method adapted to our set-up. LoAs can be nested, disjointed and overlapping and do not need to be hierarchically related or ordered (as atomic components in a molecular). This method makes the ontological commitment of a theory (in our case of a configuration of theory-method-data) explicit as follows: A configuration of theory-method-data (theory in the original Floridi scheme) comprises at least one LoA and one model. The LoA allows the theory-method-data configuration to analyse the system under investigation and to elaborate a model that identifies some properties of the system (in our set up context of intervention systems) at the given LoA. The ontological commitment of a theory can be recovered by distinguishing between a committing and a committed component, within the scheme (see Figure 3, Panel A). By making the ontological commitment of a theory-method-data configuration explicit, the method of abstraction plays a crucial role in any context in which we acquire and process information. The resulting model, then, consists of informational objects and processes.

This is the implementation problem along the Data dimension. It has to be noted (along the Methodology dimension) that in this framework, multi-methodology refers also to empirical model building in Statistics, Econometrics and Information Theory (Box and Draper, 1987; Golan, 2008; Judge and Mittelhammer, 2011). On the other hand, the Theory dimension refers to people (and organizations of people) which are involved in the research activities. How translations occur in such a way that processes of abstraction (from the local to the global) and of instantiation (from the global to the local) modify the actors involved in the translation process (Callon and Latour, 1981; Callon, 1986a,b; Latour, 1993). This is what we learn from sociology, to complete our generalized implementation problem, whose main references are summarized in Figure 2. See Figure 3 Panel B for an illustration: It highlights the translation and the relative configurations and reconfigurations of mediations originated by the movements of the instantiation and abstraction that transform the actors involved in the process. This is why it is so difficult to trace the social (Latour, 2005).
Figure 2 An Application of the framework presented in Figure 1 to address the problem of Implementation in Research Assessment.

The doubly conditional performance evaluation model and its components

We introduce now a doubly conditional performance evaluation model which is one possible model for the assessment of research coherent with the framework outlined in Figure 1. It is illustrated in Figure 4. The performance evaluation model unfolds mainly along the methodological dimension, while Theory and Data dimensions are specific to the problem content and application context.

Performance measurement in public management is a challenged subject (Johnsen, 2005). Woelert (2015) identifies a proliferation of performance indicators in a kind of technical and quantitative escalation. Lewis (2015) highlights the almost neglected consideration of the consequences of performance measurement in evaluation of public policies. As we have seen in the previous section, the output of the research activity has some features that include complexity, uncertainty and indeterminacy. Among the challenges of their assessment, there are: 1) to bring about communication and debate about assumptions, choices and uncertainties, and about the limits of scientific knowledge; 2) to allow for articulation of different types of (scientific, local, indigenous, political, moral and institutional) knowledge; 3) to provide room for a transparent negotiation among standpoints (participatory processes)(van den Hove, 2007, see p. 815 for more normative requirements for science-policy interface).

This performance evaluation model can be helpful in combining the advantages of the partial convergence indicators approach (Martin and Irvine, 1983; Martin, 1996) mainly measurability and possibility of calculating the indices of performance, with a need for a multidimensional approach to the assessment of research and its impact (Moed and Halevi, 2015).

Martin and Irvine (1983) introduce the idea of converging partial indicators approach for assessing scientific performance, based on the multidimensional nature of research and its...
outputs: “All quantitative measures of research are, at best, only partial indicators influenced partly by the magnitude of the contribution to scientific progress and partly by other factors. Nevertheless, selective and careful use of such indicators is surely better than none at all. Furthermore, the most fruitful approach is likely to involve the combined use of multiple indicators. However, because each is influenced by a number of other factors, one needs to try and control for those by matching the groups to be compared and assessed as closely as one can” (Martin, 1996, p. 351). Hence, given the partial character of indicators, it is only possible to draw reliable conclusions in those cases in which the indicators provide convergent results,
Keeping the influence of non-relevant factors low. Moed and Halevi (2015) rather than focusing on a single output dimension and underlining the need to obtain convergence among a set of different indicators in order to produce valid and useful outcomes, they aimed at proposing “a consolidated multidimensional methodological approach addressing the various user needs, interests and purposes, based on the notion that indicators designed to meet a particular objective or inform one target group may not be adequate for other purposes or target groups.” Diverse institutional missions, and different policy environments and objectives, require different assessment processes and indicators. They focus on the purpose, objectives, and policy context of research assessments, and demonstrate how these characteristics determine the methodology and metrics to be applied.

For instance, publication counts are useful instruments to help in discriminating between those staff-members who are research active and those who are not, but are of little value if research active scientists are to be compared one with another according to their research performance. Moed and Halevi (2015) introduce the concept of a meta-analysis of the units under assessment in which metrics are not used as tools to evaluate individual units, but to reach policy decisions regarding the overall objective and general setup of an assessment process. Their base underlying assumption, that we share here, is that “the future of research assessment exercises lies in the intelligent combination of metrics and peer review. A necessary condition is a thorough awareness of the potentialities and limitations of each method.”

Figure 4 provides a qualitative illustration of the main component of the doubly conditional performance evaluation model that we propose here. It is based on a combination and extension of Johnsen (2005); van den Hove (2007) and Lewis (2015). It is “doubly conditional” because the evaluation is conditioned two times: on the information we consider and on that we do not consider.

In Figure 4, the different types of arrows and the different types of boxes simply illustrate heterogeneous influences and contents respectively. The elements filled with gray represent the main items on which the conditioning could be done.

We distinguish two kinds of conditioning. Conditioning on the items reported in the bottom of the Figure 4 (policy, objectives, actors, processes and results) means to compare comparable entities, setting appropriate reference sets. We call this internal conditioning or normalization. On the other hand, conditioning on the items reported in the top of the Figure 4 means to account for heterogeneity factors that we call external conditioning or contextualization. According to this model of performance evaluation, it is all a matter of appropriate normalization and contextualization.

This model:
1. Permits the identification of the components of the analysis (in terms of theory-method-data characterization) that are excluded (what remains outside) in the specific context of the evaluation;
2. Gives interpretative value of the measure (or metrics) of research assessment calculated, that has to be considered as a residual, i.e. what remains after the consideration of the dimensions we pursued, that is due to other factors/components not accounted for;
3. Represents a step toward the democratization of the evaluation practice, able to balance the opposite views of external accountability and internal improvement (Ewell, 2009), composing contrasting trends towards competition and cooperation through cohesion.

This performance evaluation model might be helpful to identify constitutive effects of indicators (Dahler-Larsen, 2014) and perhaps also their “unintended consequences”. This model is useful for the interpretation of the results obtained from our assessment, and it is useful to identify discrepancies, and shows what the residual, our “ignorance” is. It is also helpful to identify the gap and which variables may be added to explain a part of these discrepancies. It is a kind of
contemporary revisiting and revalidation of the Leibnsteins *x-inefficiency* concept (Leibenstein 1966, 1975, 1978a,b; see also Leibenstein and Maital, 1992).

Figure 4. Another application of the framework of Figure 1: A “doubly conditional” performance evaluation model

The indicators empirically calculated are interpreted as the *residual or our ignorance* on the phenomenon and it is possible to identify the neglected aspects of the analysis carried out. The neglected components can be useful for suggesting alternative or additional dimensions of research assessment, of interest for the subjects of the assessment that are scholars, institutions and so on.

This doubly conditional performance evaluation model may be used for doing a *profiling* of the activities, sharing evaluation models, improving knowledge by learning. Within this context, *for each subject under assessment a dimension of performance along which the evaluated entity can outperform or do better than the others can be found.* The identification of the best performing dimension of each entity subjected to the evaluation is important for developing strategy for identifying and establishing sustainable and durable value creation, going beyond competitive advantages (Zenger, 2016), exploiting the existing information (Porter and Millar, 1985), and finding out for each entity its own specialty-role in the knowledge production system.

This *performance evaluation model* may be used as a *flexible evaluation tool* for “value creation” in a learning and participatory environment. It might be seen as a revisited version of Ricardo’s approach of *comparative advantages* (Ricardo, 1817) but within the context of the broader framework including theory, methodology and data dimensions introduced at the beginning of this paper.
Conclusion
In this paper we analysed the importance of designing models for the assessment of research and its impacts. We have illustrated the central role played by a comprehensive framework, which includes, in an original way, together with the Theory and Methodology dimensions, also the Data dimension. We described some limited examples of application of this framework, which refer to university rankings, the complexity of the assessment of research connected to the generalized implementation problem and the doubly conditional performance evaluation model. They show the potential of this approach and encourage to test and deepen its use in different research assessment contexts.

Acknowledgments
Financial support from the Project Sapienza Awards 2015 n. C26H15XNFS and the Project FILAS RU 2014-1186 is gratefully acknowledged.

References
Methodological Challenges for the Comparison of Results of Topic Extraction from Scientific Literature

Theresa Velden
velden@ztg.tu-berlin.de
Technische Universität Berlin, Berlin (Germany)

Abstract
This paper discusses the specific challenges of comparing topic extraction approaches in scientometrics. The discussion is inspired by experiences made in the recent topic extraction comparison exercise documented in a special issue of the journal Scientometrics (Gläser et al., 2017). It reviews the related, ongoing debate in network science about the status of ground truth in benchmarking community detection algorithms and then moves on to discuss challenges when comparing topic extraction solutions without access to a ground truth data set. The paper concludes with a proposed way forward. This paper is intended to inform the discussion at the special session on ‘Comparing Topic Extraction Approaches’ at the upcoming ISSI 2017 conference in Wuhan, China.

Conference Topic
Methods and techniques

Introduction
In a recent comparison exercise (Gläser et al. 2017, Velden et al. 2017b), several groups active in bibliometrics have conducted a systematic comparison of their respective topic extraction approaches and the results they generate. Specifically, they applied their approaches to a shared data set of bibliographic information of articles published in Astronomy and Astrophysics journals between 2003 and 2011 as indexed by the Web of Science (called the Astro Data Set). Since then, and courtesy of a license agreement provided by Clarivate Analytics, this initiative has been opened up to broad participation through the launch of a topic extraction challenge (see www.topic-challenge.info) that makes the Astro Data Set available to the community to test their approaches and conduct further comparisons. Against this background, this article discusses the various methodological challenges encountered when we seek to compare and evaluate top extraction approaches in bibliometrics in an effort to help move the field forward towards developing standards for the systematic evaluation of data analysis methods that seek to operationalize and delineate entities such as 'research topics' that are of fundamental importance to scientometric research.

For the purpose of this article, we define a topic extraction solution as a grouping of scientific papers (such as the papers in the Astro Data Set), allowing for multiple and weighted assignments of individual papers to such groups. In the simplest case, topic extraction solution is defined by a disjoint grouping of documents into clusters and fixed membership strength of documents in clusters with either 0 or 1. In more complex cases, solutions are defined as overlapping clusters and a variable strength of membership for documents within a cluster.

The motivation for comparing topic extraction approaches and solutions is to improve our systematic knowledge about methods, their appropriateness and shortcomings, specifically in the domain of scientometrics. In particular, we want to raise awareness of a problem called

1558
‘performativity fallacy’, namely the failure to consider the extent to which empirical phenomena that we observe are products of and expose features that are specific to the methods we employ (Law 2009). In the scientometrics literature, the empirical phenomenon of a topic is frequently defined ‘operationally’ as the outcome that an algorithm produces, without a careful assessment of the assumptions build into the data model and algorithm used and reflection on how those assumptions construct what is then taken as representing a topic and assessment of how topics world look differently if different choices were made in the construction process (Gläser et al. 2017). For example, after clustering a citation network of publications from a field of research, the clusters of documents we retrieve become ‘the topics’ that constitute the field. Topic extraction approaches produce the reality of topics, however due to lack of systematic investigation and comparison of our methods, we have a hard time to ascertain what properties of the topics we obtain are artifacts of our methods rather than properties of the genuine empirical manifestation of a theoretical well-defined object (‘research topic’).

To shed light on the performative nature of topic extraction algorithms, the comparison exercise mentioned above explored how much and in what ways solutions differ and how decisions that we take in the process of constructing a topic extraction solution, including data selection, data modeling, and the choice and application of an algorithm, generate solutions with distinct properties. In this article we will embed the observations made during this specific comparison exercise in a discussion of the increasing level of complexity of the challenges encountered when we seek to systematically compare and evaluate topic extraction approaches in bibliometrics.

Beyond benchmarking

At first sight, a systematic comparison of topic extraction approaches in scientometrics may seem like an obvious replication of a fairly established practice in computer science, namely to benchmark algorithms by comparing their performance when applied to the same data set. Benchmarking in computer science typically aims at evaluating the performance of a newly developed algorithm relative to existing ones with regard to speed and accuracy. For example, the development of new clustering algorithms for community detection in complex networks, routinely involves the comparison of a new community detection algorithm to other community detection algorithms (Lancichinetti et al. 2008). A popular benchmark data set for community detection algorithms that claim any relevance to social network analysis is the Zachary (1977) karate club network. This is a network of 34 nodes representing individual club members and weighted links between nodes representing the intensity of their social ties outside of the karate club. Importantly, these benchmark data sets are usually ‘labeled’, i.e. they come along with a ground truth, a specification of the true membership of each network node in a cluster stemming from direct observation rather than indirect inference. In the case of the karate club network the ground truth is based on a 3-year ethnographic observation of the club and its membership, and specifies the fission of the club membership into two separate sub networks in the course of an escalating dispute over a raise in fees for karate lessons. The accuracy of results delivered by a clustering algorithm is then measured in terms of agreement of the result with this ground truth (for which various different quantitative measures can be devised), and a popular test for community detection algorithms is whether it replicates the fission of the karate club network documented by Zachary (1977). Alternatively, one may fall back on or complement the external ground truth measures with internal quality measures that in some way assess the separation between clusters versus the internal coherence of clusters. For example see the systematic evaluation of clustering algorithms in the biomedical domain by Wiwie et al. (2015) that comprises an online evaluation service provided at https://clusteval.sdu.dk/1/mains). However it has been found that the agreement between such
internal measures and external measures of accuracy can be tenuous and what structural quality of a cluster is most relevant may well depend on the application domain, i.e. the entity that a community or cluster extracted by a method actually stands for (a scientific topic, a social group, a protein family, etc.).

The comparison exercise of topic extraction approaches in scientometrics that we discuss here differs in two important aspects. First, the approaches that we compare consist of an algorithm (typically a community detection algorithm) in combination with a bibliometric data model. This increases the complexity of the exercise, as approaches vary with regard to how the data is modeled (e.g. a direct citation network versus a similarity matrix based on word co-occurrence) and the community detection algorithm used. Second, there is no agreement on what would represent a correct topic extraction result. To the contrary, the suggestion is that when it comes to scientific topic structures there exist multiple equally valid topical perspectives onto a knowledge domain (Gläser et al. 2017). This implies that we are not looking for the one best method of topic extraction, but for guidance on what method may be best suited for a given purpose and the topical perspective that aligns with that purpose. To complicate matters further, as of now there exist no well established typology of topical perspectives nor corresponding ground truth data sets such that we are faced with the challenge to learn from direct comparisons of approaches and their results without reference to a given (set of) ground truth(s).

But let us take a step back and consider the methodological challenges of comparing and evaluating topic extraction approaches systematically, starting with comparisons that reference a ground truth data set, before discussing comparisons without reference to a ground truth data set.

**Comparison with a ground truth data set**

*The operationalization of communities in community detection algorithms*

The way a ground truth data set such as the Zachary karate club network has been used hitherto in network science took for granted that the structure of the network indeed foreshadowed and is causally linked to the eventual fission of the network that is specified by the ground truth data. Hence an algorithm that splits the network structure and separates the nodes in accordance with a given ground truth is seen as performing adequately, and the relationship between the data model (the network structure) and the ground truth is seen as unproblematic. Under these assumptions, the comparison of the results of algorithms with a given ground truth would serve to weed out algorithms that suffer from algorithmic shortcomings or biases in analyzing network structure that lead to deviations from the expected ground truth results. Aside from technical shortcomings, failures of algorithms to reproduce the ground truth communities point to a mismatch between theoretical properties of the object of study (assumed to be truthfully represented in the ground truth data) and the specific operationalization by the algorithm of communities as groupings of nodes in the network structure. A case in point is the work by Yang and Leskovec (2014, 2015) who compare communities detected in large real-world networks with ground truth communities (indicated by node labels). They suggest that many of the current community detection algorithms that allow for community overlap make the false assumption that nodes in overlap areas are weakly connected rather than strongly connected to each other. This misassumption makes those algorithms prone to misidentify overlap regions as communities in their own right.
The relationship between data model and ground truth

The question how to explain failure of community detection algorithms when compared with ground truth data sets has experienced an interesting twist recently. In a series of experiments with classical and more recent large-scale networks that have ground truth labels, Hric et al. (2014) find that none of the current community detection algorithms (overlapping and non-overlapping) performed satisfactory when compared against those ground truth data sets, leading the authors to question the relationship between data model (the given network structure) and ground truth (Hric et al. 2014). They raise the question whether the network structure adequately models the phenomenon of interest that is represented by the ground truth. They suggest that to recover the ground truth groupings one may need to enrich the data model by other than structural information and they point to respective recent efforts by computer scientists. Peel et al. (2016) go one step further, and question the readiness with which node metadata is commonly given the status of ground truth representing the study object of interest. They suggest that the node labels may refer to some other aspect or dimension of the empirical phenomenon of interest that may not bear a causal relationship with the network ties. They argue that a typical benchmark exercise simultaneously tests the performance of the algorithm as well as the correspondence of the ground truth data set to the phenomenon of interest without being able to distinguish between the two.

Applied to Zachary karate network the question is threefold: First, does the way a selected community detection algorithm operationalizes communities fit well with the theoretical model of the object of interest, e.g. the karate club network exposes features that could be exploited to discover a leader-follower partition as well as the social group fission partition, depending on algorithmic predispositions (see Fig. 1 in Peel et al 2016.) Second, does the data model capture information that is causally related to the generative process of the object of interest, so is the social fission in membership that eventually occurred have anything to do with the social ties outside the karate club between the individual club members that defined the structure of the network? In the case of the karate network, ground truth data and network structure stemmed from multi-year ethnographic observation, and the likelihood of a causal connection between data model and and object of interest is high: informed by social theory about fission in small groups it rested on the hypothesis that differential flow of information and sentiment due to the topology of the social network determine the eventual partition of the network. Third, are the ground truth labels for each club member that specify his or her membership into the old or new club after the fission appropriate to represent the object of interest? For the karate club network the ground truth data is closely tied to the object of interest, as it constitutes and describes the phenomenon to be explained, the fission of the original club membership into two factions. In other cases, especially for large networks, where ground truth labels are derived e.g. from declared group memberships in social networks or publication venues in a co-authorship network, the proposition of a strong causal link would seem to deserve further scrutiny.

Figure 1 serves to summarize this discussion. At the center of the endeavor we would expect a well articulated theoretical definition of a community (or topic) that would guide the operationalization of the respective phenomenon when modeling the data and implementing (or choosing) an algorithm to analyze this structure, as well as when generating the ground truth data set. When ground truth data is used to evaluate the performance of community detection algorithms, what is compared are the results of a data modeling and algorithmic analysis process on the one hand, with a constructed ground truth data set on the other hand. Too often in the community detection literature, the relationship of the theoretical concept of study to the data model and the ground truth is taken as unproblematic and remains exempt from scrutiny. However, in any of these steps,
problematic decisions can be taken that veer the results of the community detection approach and the results suggested by the benchmark away from the actual phenomenon of scientific interest. So what we can take away from this discussion is that when we interpret the benchmarking results of a community or topic detection algorithm vis a vis a ground truth data set, we need to exercise extreme caution in drawing conclusions whether or not an approach is suitable to detect the kind of objects we are interested in.

Fig. 1: Diagram depicting the various 'performative' steps undertaken, before a community detection result (or topic extraction result) is compared to and evaluated against a ground truth data set. In any of these steps decisions are taken that need to scrutinized with regard to their correspondence to the theoretical properties of the object of study.

Comparison without a ground truth data set
As mentioned above, there is no agreement on a single topical perspective that would prescribe a correct partition of a set of documents in the scientific literature. Topical structures are seen as multidimensional, as commonalities between research contributions may arise along different dimensions such as research objects versus research methods (see for example discussions in Wen et al 2017, Glänzel & Thijs 2017, Palchykov et al. 2016, and Newman & Girvan 2002) however as of now no consolidated theoretical knowledge exists that would allow us to specify these different topical perspectives. For the time being this renders the respective multiple ground truths
inaccessible. Under these circumstances, the question of how to conduct a comparison without reference to a ground truth and what to learn from it about the suitability of different becomes a research question in its own right.

We have seen scientometricians take different approaches to this question, as exemplified in the recent literature by Klavans & Boyack (2016), Subelj et al. (2016), and Velden et al. (2017).

**Ad-hoc yardstick**

Ad-hoc yardsticks take on the role of a ground truth data set. Where classical ground truth data sets ideally hail from independent observation, ad-hoc yardsticks are constructed on the spot, from the available data. For example, Klavans & Boyack (2016) investigate the influence of different citation based data models (co-citation versus direct citation versus bibliographic coupling) on topic detection results using the same community detection algorithm. As a yardstick for comparing and assessing the ‘accuracy’ of the different solutions they use overlap with the grouping of references by review articles (articles with > 100 references), and conclude that the direct citation data model performs best with regard to this yardstick. What Klavans’s and Boyack’s yardstick definition has going for it is that it is informed by domain knowledge developed over a career of doing scientometric research on the science system. They distinguish between scientific topic perspectives with different characteristics (research fronts, a topic of an earlier comparative study using a different yardstick (Boyack & Klavans, 2010), versus taxonomic topic classification) and seek to design their yard sticks to match those distinct perspectives. Whether such a yardstick would be appropriate to compare and evaluate a wider range of data models (including those that do not use citation as a feature) is doubtful, since the yardstick seems to be too closely related to citation based data models and is likely biased towards it.

**Statistical properties**

Subelj et al. (2016) compare a large range of different clustering algorithms by applying them to direct citation networks from different domains. The specific purpose that results would be used for is not elaborated beyond a general description of the task as “grouping scientific publications into clusters.“ In their comparison they focus on structural properties of solutions, complemented by their own expert review for one of the domains they can have expertise in (scientometrics). They find that the different approaches can be subdivided into a small set of families yielding similar results based on a metric of pairwise clustering differences. Those different families of algorithms tend to generate distinct size distributions which in turn a related to some distinct properties of linking patterns found with regard to internal and external linking, as well as further metrics relating to range of cluster sizes and cluster diameters. The authors suggest that reviewing such statistical properties provides insights that allow to discriminate between algorithms more suitable for the task.

---

1 Interestingly, the lack of ground truth data implies that no preferred level of resolution is defined anymore. As a community we have not yet developed a consistent way for how to deal with and think about different resolution levels in the extraction of scientific topics from the literature and their comparison. Velden et al. (2017b) attempted comparisons across widely differing resolution levels (ranging across approaches from 11 to 555 topics retrieved) that were in several cases decided ad-hoc rather than based on some theoretical justification. Subelj et al. (2016) in their comparison of community detection algorithms simply accepted the default setting provided by most algorithms, also resulting in a variety of different resolution levels to compare across. For comparison purposes one might tune algorithms to deliver very similar numbers of topics, and scan a range of such levels. Or one might search for resolution levels that seem the most natural for a given combination of data model and algorithm, based on some internal quality criterion, as discussed by Wiwie et al (2016).

2 Whether many current ground data sets for large data sets (such as social networks where the data model are friendship links between users and ground truth data are simply the group memberships declared by users using a respective feature of the same social network site) actually fulfill this requirement can be debated.
or less suitable. They develop a number of intuitive and practical criteria such as robustness of results, avoidance of excessive run times, a minimal proportion of publications being assigned to clusters of insignificant size, and a preference for cluster sizes ideally to not vary by more than one order of magnitude. However, they suggest that the ultimate criterion should be expert review of the topics obtained. Based on these criteria along with a favorable assessment of topic content by their own expert review, they conclude that infomap (or alternatively Metimap for very large networks) is the most suitable method to use on direct citation networks.

Embracing variation

Rather than looking to determine the most accurate or suitable approach by one measure or another, two recent works embrace variation and start from the assumption that a multiplicity of legitimate topical perspectives exist. Wen et al. (2017) triangulate three different bibliometric mappings of topics in the field of water research to demonstrate the complementarity of information gained by different approaches, journal to journal citation, author keyword overlap, and title word + reference overlap. They suggest that those complementary methods identify fields, topics and methods used within the field of water research.

Velden et al. (2017b) come up with various ways to determine and visualize differences between the topical structures produced by various approaches when applied to the same data set (Astro Data Set) and attempt to determine how choices made in the design of an approach produce specific outcomes. The latter goal of tracing back features of solutions to specific decisions in the design of an approach is challenging, as the multiplicity of data models and algorithms generates a large space configurations, and one would ideally compare solutions by varying only one aspect at a time, and addition scan a range of different levels of resolution to facilitate direct comparison. Based on their sample of approaches, Velden et al. (2017b) were dealing with a 5 by 5 matrix of potential combinations of data models and algorithms (see Figure 2), however only had resources to investigate and compare 8 out of the 25 possible combinations. Note that this matrix does not consider investigations of the same approach at various levels of resolution. Resolution levels varied widely across the sample, making direct comparisons problematic for many pairs of solutions.

Velden et al. find that all solutions in their sample agree in grouping together certain sets of documents, suggesting the existence of topical ‘cores’ of very densely connected documents by various types of community measures and represented by various data models. Solutions differ across a range of structural features, such as concentration of topic sizes and coverage of the data set, the latter ranging between 93% and 100%. Interestingly, one approach (Koopman et al. 2017) produces a major topic with a large proportion coming from those remaining 7% of publications missing as a distinct topic from the solutions delivered by those latter approaches.

These are still rather formal structural properties, observed without reference to specific topic contents or interpretation of the topical structure of the Astro Data Set that the different solutions construct. The authors further employ a number of visualization and labeling techniques (Koopman & Wang 2017, Velden et al. 2017) to explore the topical structures generated by the various approaches. Based on their analysis, three distinct perspectives onto the Astronomy and Astrophysics field can be distinguished:
**Fig. 2:** Potential configuration space of approaches compared in original topic extraction exercise using the Astro Data Set. Of 25 possible configurations, only 8 were compared. The number of configurations to compare is multiplied when systematically scanning different levels of resolution (number of topics).

**External:** This perspective is generated by considering the linking of documents in the Astro Data Set into the wider literature (Boyack, 2017). It is a view onto the field when considering the ‘external pull’ by documents published in other disciplines. It depicts the field as highly concentrated within some core domains (Gravitation and Cosmology, Astrophysics, and Solar Physics) where almost all documents combine into one or two large topics, plus other domains (such as Planetary Science or Astroparticle Physics) that are characterized by a wide scatter of documents into many smaller topics. This topical perspective accentuates interdisciplinary links and identifies subdomains within the field of Astronomy and Astrophysics with strong links to other disciplines.

**Internal & Global:** This perspective is internal, as seen from within the field as it only considers information within the Astro Data set. The results of four of the eight approaches in the sample (by Velden et al. 2017, van Eck & Waltman 2017, and Koopman et al. 2017) show great similarity although they use a set of different data models and algorithms. Their representations as topic affinity networks map the field as an elongated structure, organized by research objects from large-scale objects distant in space-time (cosmology) to objects in the local neighborhood of Earth (solar physics, planetary science, space science). The core of the field (with research on galaxies and stars) is resolved at a much higher level of granularity than by the external perspective.

**Internal & Local:** This perspective is generated using a community detection algorithm that produces overlapping communities and only evaluates the local neighborhood of a seed of documents, i.e. not taking the complete network information of documents in the Astro Data set into account (Havemann et al. 2017). It produces topics at a large variety of sizes. As Velden et al observe, they tend to reproduce the main topics found by the internal and global approaches in the sample, but also some distinct topics that are not identified by the latter approaches. By allowing for overlap, this approach may be more suitable to detect ‘bridging’ or ‘emerging’ topics at the interfaces of larger established topics that may get suppressed by approaches that only allow for disjunct topics constructed from a global analysis of network structure.

<table>
<thead>
<tr>
<th>DATA MODEL</th>
<th>Direct citation</th>
<th>Bibliographic coupling</th>
<th>Hybrid (bc + NLP)</th>
<th>Semantic Matrix</th>
<th>Global direct citation map</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ALGORITHM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Informap</td>
<td>✔</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>SLMA</td>
<td>✔</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>✔</td>
</tr>
<tr>
<td>Louvain</td>
<td>−</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>−</td>
</tr>
<tr>
<td>Memetic</td>
<td>✔</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>k-means</td>
<td>−</td>
<td>−</td>
<td>✔</td>
<td>−</td>
<td>−</td>
</tr>
</tbody>
</table>

**Levels of resolution**

<table>
<thead>
<tr>
<th><strong>A</strong></th>
<th><strong>L</strong></th>
<th><strong>G</strong></th>
<th><strong>O</strong></th>
<th><strong>R</strong></th>
<th><strong>T</strong></th>
<th><strong>H</strong></th>
<th><strong>M</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>✔</td>
<td>−</td>
<td>−</td>
</tr>
</tbody>
</table>

1565
These distinctions underline how the suitability of approaches is dependent on purpose and intended use of the results of a topic extraction. They are first indication of what one might consider as topical perspectives, but fall short of a well-founded theoretical typology.

The Way Forward
It seems obvious at this point, that topic extraction approaches and their results need to be considered relative to their purpose and intended use, and in particular relative to the topical perspective that an approach is intended to evoke from the data. So one important line of research is to develop a typology of topical perspectives and corresponding theoretical definitions of topics that would need to inform their operationalization in form of data models and algorithms. One way would be to design an empirical study to engage with practitioners (scientists within the field, from neighboring fields, consultants, funding agencies, information services) to try disentangle distinct topical perspectives that are relevant in specific contexts for distinct purposes. How do these topical perspectives differ, and what are the implications for the theoretical conception of topics and their operationalization in the form of data models and algorithms?

Then, for a given purpose and topical perspective of interest, we would like to compare and evaluate different approaches. What is a good modeling and algorithmic approach to construct the relevant topical structure given theoretical insights into the nature of the topical perspective and the implied requirements for their operationalization? How can we resolve the problem that many popular data models (such as citation based models) likely mix different topical perspectives, as highlighted by Palchykov et al. (2016)? What can serve as a yardstick or ground truth to compare and assess the suitability of competing approaches for the respective purpose? How do we map and compare results that deliver overlapping topics, given that many of the standard metrics and mapping techniques assume a disjoint construction of topics?

At this point, I would like to highlight the opportunity to join the topic extraction challenge (www.topic-challenge.info) that makes available the Astro Data Set: to take advantage of a shared data set to learn about your own approach and how the solutions it produces compares to the solutions of other approaches; to participate in the discussion of what are relevant alternative topical perspectives, what could a yardstick to represent a ground truth look like in each case, and what are useful tools for the comparison of topical structures? The originators of the topic extraction challenge (Boyack et al. 2017) hope to establish the Astro Data set as a benchmark data set to test approaches and to grow the collection of solutions into a valuable open resource for further study as we hope to generate interest in improving collectively our systematic knowledge about methods, their suitability, strengths and limitations.

While it may seem as if we are going in a circle, I suggest we are making progress by having advanced our understanding of the various layers of performativity of topic extraction approaches, and by having identified avenues for further progress.

"We have learned a lot, Siddhartha, there is still much to learn. We are not going around in circles, we are moving up, the circle is a spiral, we have already ascended many a level."
(Hermann Hesse, Siddhartha, Chapter 2, 1922)
References


Study on the Method of Detecting the Research Fronts in the View of Topic Model

Feng Jia  Zhang Yunqiu
m18844501520@163.com
yunqiu@jlu.edu.cn
JiLin University, ChangChun (China)

Abstract
We have proposed a method of detecting research fronts based on LDA. LDA can realize the extraction of topics from corpus accurately. Many current methods of detecting research fronts are lack of the indicators to identify research fronts. For the problem, we construct the indicators-strength and novelty. In the empirical study, we detect the research fronts of medical informatics, and the results shows that there are 19 research front, beside 17 of them have been approved by experts.

Conference Topic
Research Fronts and Emerging Issues
Indicators
Knowledge discovery and data mining

Introduction
Research fronts present the focus and difficulty domain of scientific research. Obtaining the research fronts timely and accurately is of great significance for the country, institutions and researchers. Countries all around world give high attention on the study of the research fronts. Research fronts is critical with future of the development of scientific and technology, hence identifying the research fronts timely and accurately is of great importance. What’s more, academic resources are increasing day by day. Researchers are faced with great challenges in understanding the development trend of the field and obtaining research fronts. In addition, there are still some problems of research fronts identification (Leydesdorff, 2011).

1.1 Literature Review of Research Fronts
Price (1965) proposed the concept of research fronts. Afterwards researchers conduct abundant researches on it. Research defined the meaning of research fronts in different perspectives, and it can be classified into four categories: ①define as a collection of high cited papers, eg. Price (1965); ②define as a collection of co-cited papers, eg. Garfield (1994); ③define as a collection of bibliographic coupling papers, eg. Morris (2003); ④ defined as an emergent and transient grouping of concepts and underlying research issues, eg. Chen (2006).

There are various methods of detecting research fronts, which can be summarized from the qualitative and quantitative perspective. In the qualitative, literature review and brain storm are commonly used and authoritative methods. However, researchers are with different capabilities, so the result is obviously subjective and inconsistent. In the quantitative, it can be combed into citation analysis method and content analysis method, which are the focus of method studies.

With ample research results, citation analysis method started early, which can be concluded into co-citation analysis, bibliographic coupling, and direct citation. Recent years, co-citation analysis and bibliographic coupling are widely applied in detecting research fronts. Lee (2015) identifying the research fronts in Korean library and information science by document co-citation analysis. Liu (2016) conducted the study which applies a network clustering method to group the literature through a citation network established from the data envelopment analysis (DEA) literature over the period 2000 to 2014. Each research front is
then examined with key-route main path analysis to uncover the elements in its core. Due to the citation delay and it can’t reveal the content of paper, many researchers attempt to detect research fronts by content analysis method.

Content analysis method includes term frequency analysis, burst term analysis, and co-word analysis. The objects of term frequency analysis are keywords or subject headings mostly. Chu (2014) and Qin (2016) detected the research fronts based on term frequency analysis. However, this method has a drawback of the threshold need to be set manually, which may lead to the subjective results. In 2006, CM Chen embed Kleinberg’s (2003) burst detection algorithm into CiteSpace II, which is adapted to identify emergent research-front concepts. Compared with the terms analysis, co-word analysis emphatically measures the relationship between terms. Han (2015) adopted co-word analysis method to detect the hotspot domains and knowledge base in the research of special education teachers. The keywords, subject headings, burst items, and couples of co-words are condenses of the content of paper, which can represent the core meanings and research directions. Muñoz-Écija (2017) identified their origins of nanoscience and nanotechnology (NST) and the seminal papers through direct author citation of these works, and extracted from the titles and abstracts through word co-occurrence, the main lines of research were identified.

Through the literature review, there are two main problems in the research of front identification:

1. It’s lack of recognizing indexes of research fronts.
   The existing methods of detecting research fronts depend on the accumulation of terms or citation, though it’s lack of recognizing indexes.

2. The existing methods neglect the semantic information between texts.
   Citation analysis and content analysis method can’t detect the research fronts semantically, and they neglect the semantic information between texts.

With the development of text mining technology and semantic analysis, it is the research focus and future trends that detecting the research fronts accurately and semantically.

1.2 Literature Review of Topic Model

In the scientific researches, topics generally mean the main content in the paper. Under the topic model, topics are represented by a series of probability allocation. Topic model is a latent modelling method which can extract the topics from text. What’s more, the topics can be used to analyse the corpus preliminary. Up to date, many researches demonstrate that topic model can extract topic from massive text in the era of big time. In addition, one of the most important reasons of the prevalent is that the topic model is not so complicated and more explainable compared to other text mining methods.

Among the massive topic models, Latent Dirichlet Allocation(LDA) is a complete Bayesian model with a solid theoretical foundation. Due to the characters of flexible and expendable, it is widely used in emotional analysis, social media analysis, academic corpus analysis, and network structure data. The paper proposed the method of detecting research fronts based on LDA, and emphasized the application in the academic corpus analysis.

The earliest used in the academic corpus analysis of LDA is Grittith (2004). He used LDA model and showed that the extracted topics capture meaningful structure in the data, consistent with the class designations provided by the authors of the articles, and outline further applications of this analysis, including identifying “hot topics” by examining temporal dynamics and tagging extracts to illustrate semantic content. Hall (2008) apply unsupervised topic modelling to the ACL Anthology to analyse historical trends in the field of Computational Linguistics from 1978 to 2006. He induced topic clusters using LDA, and examined the strength of each topic over time. Zhou (2006) proposed a method for discovering the dependency relationships between the topics of documents shared in social
networks using the latent social interactions and trained a LDA model over our entire sample collection of CiteSeer, attempting to answer the question: given a seemingly new topic, from where does this topic evolve? Gideon (2006) extends journal bibliometric to topics (Citation Count, Impact Factor, Diffusion, Half-Life) and introduces three new topic impact measures: Topical Diversity, Topical Transfer, and Topical Precedence.

In the domestic, researchers mainly apply LDA to detect emerging topics. Fan (2014) attempted to detect the emerging topics from documents based on LDA. Wu (2012) proposed ATNLDA model to mine the topics evolution, which based on the methodology of co-occurrence and cluster. Lee (2015) used LDA and sparse representation classifier to categorize new topics. Ye and Leng (2013) combined the LDA and co-occurrence.

The studies of researchers demonstrate that LDA performed well in the topic extraction based on the academic corpus. The paper proposed the method of detecting the research fronts which is extracting the topics by LDA and identifies the front topics by research front indicators. We conduct empirical study with the example of medical informatics discipline, and access the results by experts.

2 Methodology

2.1 LDA

LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modelled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modelling, the topic probabilities provide an explicit representation of a document.

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

LDA assumes the following generative process for each document \( \mathbf{w} \) in a corpus \( D \):

1. Choose \( N \sim \text{Poisson}(\xi) \).
2. Choose \( \theta \sim \text{Dir}(\alpha) \).
3. For each of the \( N \) words \( w_n \):
   
   a) Choose a topic \( z_n \sim \text{Multinomial}(\theta) \).
   
   b) Choose a word \( w_n \) from \( p(w_n | z_n, \beta) \), a multinomial probability conditioned on the topic \( z_n \).

The LDA model is represented as a probabilistic graphical model in Figure 1. As the figure makes clear, there are three levels to the LDA representation. The parameters \( \alpha \) and \( \beta \) are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. The variables \( \theta_d \) are document-level variables, sampled once per document. Finally, the variables \( z_{dn} \) and \( w_{dn} \) are word-level variables and are sampled once for each word in each document.

![Figure 1: Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.](image-url)
We can view LDA as a dimensionality reduction technique, but with proper underlying generative probabilistic semantics that make sense for the type of data that it models.

2.2 Indicators of Identifying the Research Fronts

LDA can extract the latent topics from big amount corpus which contains many research front topics. So how can we detect them accurately and quickly? We try to structure the indicators to identify them. The paper proposed the strength and novelty of topics as the indicators. Strength can reflect the high attention by researchers, and novelty can reflect the latest studies. To illustrate the strength and novelty of topics, we introduce the concept of topic firstly.

Topic is a series of items \(w_i\) with its weight \(p(w_i|z)\) which can reveal the semantic meaning of topic. It represents in means of vector:

\[
\text{z} = ((w_1, p(w_1|z)), (w_2, p(w_2|z)), \ldots (w_n, p(w_n|z))) \quad \text{Formula (1)}
\]

Figure 2 is the text expression of topic:

\[
\begin{pmatrix}
T_{t_1} & T_{t_2} & \ldots & T_{t_n} \\
\begin{pmatrix} p_{t_1} & p_{t_2} & \ldots & p_{t_{n+1}} & p_{t_{t_n}} \end{pmatrix} & \begin{pmatrix} p_{t_2} & p_{t_2} & \ldots & p_{t_{2n+1}} & p_{t_{2n}} \end{pmatrix} \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
\begin{pmatrix} p_{t_{k-1}} & p_{t_{k-1}} & \ldots & p_{t_{k(n-1)}} & p_{t_{k-1n}} \end{pmatrix} & \begin{pmatrix} p_{t_k} & p_{t_k} & \ldots & p_{t_{kn-1}} & p_{t_{kn}} \end{pmatrix}
\end{pmatrix}
\]

Figure 2: Text expression of topic.

Strength of topic is the indicator to access the topics whether it can be hot topics. In general, strength is the ratio of the overall document amounts to the amounts of one topic involves. The formula is following:

\[
\theta_j = \frac{\Sigma_d \theta_j^{(d)}}{M} \quad \text{Formula (2)}
\]

\(\theta_j\) is the value of strength of one some topic. \(M\) represents the total amount of documents. \(\theta_j^{(d)}\) represents the weight of topic \(j\) in the document \(d\). The weights are more high, the more important. Hence, we can calculate the strength of the topic and plot the trends to illustrate the topic.

Novelty of topics can reflect the average publication time. If one topic publishes more recently, the higher degree of the novelty. Research fronts are the latest scientific findings and research issues, so they are more novel than other topics. In this paper, we calculate the novelty by counting the average publish time.

2.3 Workflow of Detection of Research Fronts

Figure 3 is the workflow of the detecting of research fronts based on LDA. It includes three periods, which are data pre-processing, extracting the topics based on LDA, and identifying the research fronts.

In the data pre-processing period, we collect data by specific retrieval strategy from literature database. Then remove the duplicate records, delete incomplete records, and extract “Abstract” field. Abstracts are the source of corpus. At last, remove stopwords and do lemmatization, so the corpus can be constructed.

In the period of extracting the topics based on LDA, it focuses on the three parameters estimation: the Dirichlet hyperparameters \(\alpha\) and \(\beta\), and the number of topics \(K\). In this paper, \(K\) is set by the Perplexity of topic model, and when the value of Perplexity is smallest, \(K\) is
the optimal value. For the hyperparameters $\alpha$ and $\beta$, they are set as $\alpha = 50/K$, $\beta = 200/W$ (W is number of words in the vocabulary).

In the period of identifying the research fronts, it needs to calculate the strength and novelty of each topic. When the strength and novelty of one topic reach the threshold, it can be identified as research front.

Figure 3: The workflow of the detecting of research fronts based on LDA.

3 Empirical Study

3.1 Data Pre-processing

We choose Web of Science$^{\text{TM}}$ as the data source and adopt “subject categories” to retrieval the documents of the discipline of medical informatics. Web of Science$^{\text{TM}}$ contains 252 subject categories, and “medical informatics” is one of them. From 2007 to 2016, there are 35981 records under the subject categories of medical informatics (Retrieved on January 3, 2017).

Table 1. Bibliography of Document Types

<table>
<thead>
<tr>
<th>Document Types</th>
<th>Number</th>
<th>Title</th>
<th>Keywords</th>
<th>Abstract</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article</td>
<td>21940</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Biographical-Item</td>
<td>60</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Book Chapter</td>
<td>23</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Book Review</td>
<td>17</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Correction</td>
<td>205</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Editorial Material</td>
<td>1296</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Letter</td>
<td>442</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Meeting Abstract</td>
<td>64</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Proceedings Paper</td>
<td>11761</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Reprint</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Review</td>
<td>900</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

In this paper, the corpus need to include title, keyword, abstract, and year. Therefore, we adopt the Article, Proceedings Paper, Review, and Reprint, which amounts 33870 records and accounts for 94.13% of total number.
3.2 Results

We perform parameter estimation using collapsed Gibbs sampling (Griffiths and Steyvers, 2004). This study uses the R language to achieve topic extraction based on LDA. Figure 4 is K selection results, showing the Perplexity of the corpus for different settings of the number of topics. With the increase of number of topic, the Perplexity of the model gradually decreases. When the number of topic is 62, the model has the lowest Perplexity, which is the best fitting result. Furthermore,

![Figure 4: Trendline of Perplexity](image)

After extracting 62 topics, we calculate the strength of topic, the result is shown in figure 5.

![Figure 5: strength of topics](image)

In figure 5, the height of bar represents the value of strength, and the dotted line is the average of strength of 62 topics. The average of strength is 2.17. Topic 4 is the “Medical text knowledge extraction” with the highest strength. Topic 1, 3, 4, 9, 19, 21, 31, 41, 48, 54, and 57 are with higher strength. There are 22 topics which strength is higher than average.

In the aspect of novelty of topics, we adopt boxplot to visualize. In figure 5.5, the boxplot contains one rectangle, dotted line, and the tow borderlines. The spacing between the different parts of the box indicate the degree of dispersion spread and skewness in the data. In addition to the points themselves, they allow one to visually estimate various L-estimators, notably the interquartile range, midhinge, range, mid-range, and trimean.
To calculate the novelty of the 22 topics which strength is higher than average, figure 6 shows that the average public time of 22 topics is 2012. The public time of topic 3, 8, 19, 22, 31, 34, 53, 54, and 58 is 2013. In the view of novelty, topic 3, 4, 8, 19, 20, and 34 is more novelty than others, which active around in 2015.

Combine the strength and novelty, we set the research fronts as its strength is higher than average and public time is 5 recent years. Hence the paper identifies 19 research fronts in medical informatics shown in table 2. Due to the limited space, we cannot present all the automatically extracted top words for all topics. Instead, we manually tag all the topics with labels by summarizing the ranked words.

**Table 2 : the list of research fronts.**

<table>
<thead>
<tr>
<th>NO.</th>
<th>Label</th>
<th>Strength</th>
<th>Novelty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tumor image analysis</td>
<td>3.58</td>
<td>2011.74</td>
</tr>
<tr>
<td>3</td>
<td>Application of Data Mining Algorithm in Medicine</td>
<td>3.62</td>
<td>2012.56</td>
</tr>
<tr>
<td>4</td>
<td>Medical text knowledge extraction</td>
<td>4.83</td>
<td>2012.30</td>
</tr>
<tr>
<td>8</td>
<td>Health medical application</td>
<td>2.99</td>
<td>2012.46</td>
</tr>
<tr>
<td>9</td>
<td>Community Health Service</td>
<td>3.27</td>
<td>2011.92</td>
</tr>
<tr>
<td>17</td>
<td>clinical decision support</td>
<td>2.61</td>
<td>2011.58</td>
</tr>
<tr>
<td>19</td>
<td>New medical model based on network and computer</td>
<td>3.45</td>
<td>2012.56</td>
</tr>
<tr>
<td>21</td>
<td>Disease diagnosis system and disease classification method</td>
<td>3.27</td>
<td>2012.06</td>
</tr>
<tr>
<td>22</td>
<td>Development and Application of Medical System and Software</td>
<td>2.65</td>
<td>2012.32</td>
</tr>
<tr>
<td>24</td>
<td>Research on Integration of Medical System and Medical Data</td>
<td>2.81</td>
<td>2011.84</td>
</tr>
<tr>
<td>31</td>
<td>Evaluation of Health Information System</td>
<td>3.27</td>
<td>2011.98</td>
</tr>
<tr>
<td>34</td>
<td>Disease Survival Model</td>
<td>2.75</td>
<td>2012.38</td>
</tr>
<tr>
<td>41</td>
<td>Medical Informatics Method and Technology</td>
<td>4.16</td>
<td>2011.62</td>
</tr>
<tr>
<td>45</td>
<td>Electronic medical records</td>
<td>2.64</td>
<td>2011.50</td>
</tr>
<tr>
<td>46</td>
<td>Disease risk prediction</td>
<td>2.55</td>
<td>2011.92</td>
</tr>
</tbody>
</table>
To access the results, the paper adopts expert consult. We invite 7 experts in medical informatics to score the result by Likter five scores. The experts are mainly professors who engaged in medical informatics research many years, respectively from Peking University, JiLin University, ZhongNan University, China Medical University, and Institute of Medical Information, CAMS&PUMC. The result of accessing shows that 17 research fronts have been approved by experts.

4 Conclusion

We have proposed a method of detecting research fronts based on LDA. LDA can realize the extraction of topics from corpus accurately. Many current methods of detecting research fronts are lack of the indicators to identify research fronts. For the problem, we construct the indicators-strength and novelty. In the empirical study, we detect the research fronts of medical informatics, and the results shows that there are 19 research fronts, beside 17 of them have been approved by experts.

However, this method has some limitation. One is the threshold of the strength and novelty, which settings need to be manually involved. Anther limitation is that the corpus is only form literature database. Granted, using one or another data source may condition the results of the study. Database coverage is determinant for obtaining more or less comprehensive results. Due to the research fronts exist in many carriers, such as patent, tender and even scientific blog on the web. Future studies may therefore lead us to integrate multiple text sources to detect research fronts, so as to arrive at comparative results. In addition, some future research based on the method could be incorporated and some potential applications.

Reference


Integrating domain ontologies with topic modelling for mapping societal problems

Montserrat Estañol1,2, Francesco Massucci1, Alessandro Mosca1 and Ismael Ràfols3

1 {montserrat.estanyol | francesco.massucci | a.mosca}@sirisacademic.com
SIRIS Lab, Research Division of SIRIS Academic, Barcelona (Spain)
2 Departament d’Enginyeria de Serveis i Sistemes d’Informació, UPC, Barcelona (Spain)
3 i.rafols@ingenio.upv.es
Ingenio (CSIC-UPV), Universitat Politècnica de València, València (Spain)

Abstract
This study presents an attempt to systematically integrate domain knowledge into the algorithmic classification of large text corpora. Taking the societal problem of ‘obesity’ as case study, we build an ontology for obesity and integrate this ontology into topic modelling algorithms. The idea is that for the purpose of mapping a given societal issue with a large collection of scientific texts, the combination of a domain ontology and topic modelling may facilitate the tasks of delineating the relevant corpus, classifying it into topics and interpreting the resulting topics. Preliminary results show that the topics obtained adding an ontology to topic modelling are more meaningful. Considering the need of aligning the research with societal needs, our work goes into the direction of providing a tool to map the topics research outputs in mapping categories that are easier to interpret from a user’s perspective.

Introduction
The issue of mapping science and having a bird’s eye view of scientific production is one of the core problems of scientometrics (e.g. Börner, Chaomei & Boyack, 2003). Yet the increasing policy demands for aligning scientific endeavors with societal needs (Cassi et al. 2017) requires methods that allow policymakers to analyse research corpora with a societal problem perspective (Wallace & Rafols, 2015).

Recently, several techniques have been proposed to coarse grain the outputs of scientific production, by obtaining both maps of words (Van Eck & Waltman, 2010) and by using statistical methods to retrieve topics in large datasets (Griffiths & Steyvers, 2004). Albeit useful, these methods do not use available human knowledge in their algorithms. It is therefore poorly understood how their results differ from expert-informed classifications of scientific documents into clusters of research topics.

In the last years, preliminary steps have been made towards the integration of semantic technologies (e.g. human-based ontologies) and algorithmic clustering (e.g. machine learning methods) for the classification of large datasets (Andrzejewski, Zhu & Craven, 2009; Chhatbar, 2010). However, the applicability of the aforementioned methods is still limited, mainly because of three factors: i) domain knowledge is not usually embedded in a systematic way, so that expert knowledge is enforced locally only for a subset of concepts in the corpora; ii) models developed so far assume in general that there exist only one ‘best’ representation of knowledge, without taking into account that experts with diverse background may have different descriptions of the same domain; and finally iii.) the performance of those methods against the array of possible unsupervised alternatives is yet to be proven.

A representation of knowledge that aims at capturing a comprehensive view of scientific production must leverage on plural representations of a scientific domain, which means taking into account a variety of views. Comprehensiveness and pluralism become especially regarding
topic mapping tools for policy, given that interpretation of knowledge landscapes are dependent on the framing of the topic.

In this case study of obesity, the challenges above are tackled by, firstly, gathering the knowledge of heterogeneous experts through direct interviews; secondly, by formalising the knowledge from interviews into an ontology centred on the concept of ‘obesity’; and finally, by embedding such ontology into topic modelling algorithms, in order to produce an automated topic mapping.

Methods
The present work aims at incorporating a domain ontology into topic modelling (TM) algorithms, for the case of ‘obesity’. Let us first introduce the concepts of ontology and TM.

An ontology is a formal description of the concepts in a certain domain (where a domain is the given fragment of reality that the ontology aims to represent), the relationships between them and the properties of each of the concepts. An ontology also allows representing particular instances of a concept (i.e. individual cases??) and the relationships between them. For example, in the obesity ontology, Disease is a concept, and Obesity is an instance of Disease. A disease may be treated following a certain treatment: Treatment is another concept in the ontology, related to the Disease concept, through relationship canBeTreated. The ontology also allows us to represent that Obesity may be treated by following a VeryLowCalorieDiet, where VeryLowCalorieDiet is an instance of Treatment.

However, ontologies go further than just carrying out the descriptions listed above. On the one hand, given an ontology, it is possible to infer new knowledge by using a reasoner. On the other hand, it is possible to define axioms (i.e. restrictions) in the manner how the concepts and relationships relate to each other, so as to obtain more accurate representations. An ontology can be formalised and visualised as a graph, where nodes are concepts and edges are the logical relationships that link those concepts.

Topic Modelling (TM) is a field of machine learning that aims at ‘discovering’ the unknown topics to which a collection of texts belongs to (Steyvers & Griffiths, 2007). Generally, the topics are stacks of tightly related words that co-appear consistently in the observed texts. Several techniques have been proposed to detect the topics, but all share the idea that the texts were produced by a stochastic generative model, and aim at recovering the parameters of the model that best explain the observed data. In our work, we use Latent Dirichlet Allocation (LDA) (Blei, Ng & Jordan, 2003), which assumes a text generative model that has a Dirichlet distribution for the probability of linking topics to documents and words to topics.

The steps outlined below describe the tasks of building an ontology dealing with the theme of obesity, embedding it into LDA and retrieving domain-knowledge informed topics.

Step 1: Identification of the corpus
We used in our study the obesity corpus identified in Cassi et al. (2017) which was delineated with the following procedure. Firstly, publications with MeSH term matching the search obes* in MEDLINE/PubMed during the 2000-2013 period were retrieved. Secondly, the same search directly on the WoS Core Collection was performed. Finally, from all micro-clusters based on direct-citation paper clustering of all WoS for 2000-2013, those clusters where at least a 10% of papers tagged as obesity (according to MeSH or WoS Topic) were retained.
Step 2: Creation of an ontology around the concept of Obesity

To create the obesity ontology, two main sources were used. On the one hand, we used the information obtained by checking relevant literature, ranging from Wikipedia to scientific publications, including medical websites and World Health Organization (WHO) documents. On the other hand, we interviewed four experts to obtain their specific input on the topic. These experts came from four different fields of expertise: nutrition, medicine, biology and health policy.

All the interviewees agreed that obesity is a multifactorial disease and that it needs to be tackled with diverse approaches. Therefore, the ontology is not limited to the medical perspective: it has been informed by articles on marketing and its impact on food consumption, lifestyle, psychological, and nutritional perspectives, among others, in order to capture a comprehensive and plural view on. The graph of the resulting ontology is shown in Figure 1.

Step 3: Integration of the domain ontology with topic modelling

Previous works have focused on the effort of integrating some expert knowledge onto topic modelling. Logic LDA for instance is aimed at incorporating first order logic into topic modelling based on LDA (Andrzejewski, Zhu, Craven, & Recht, 2011). However, the task of injecting a whole ontology onto some topic modelling algorithm remains unaccomplished so far. Here, we address this challenge by decomposing the ontology built in Step 2 into modules that are fed to Logic LDA in the form of its native rules (i.e., . . . ), by adopting the following approach. As outlined previously, an ontology can be encoded in a graph, where nodes are
concepts and links are logical relations among those concepts. In turn, the resulting graph can be analyzed by means of the tools of network theory, such as, for instance, community detection algorithms. We thus partitioned the ontology graph into communities of closely related concepts by using the Infomap clustering algorithm (Rosvall & Bergstrom, 2008). We then used the resulting clusters to build rules aimed at guiding LogicLDA topic discovery. Specifically, we retrieved 33 clusters of highly connected concepts and, for each of these clusters, we enforced a LogicLDA ‘seed’ rule that enhanced the retrieval into the same topic of groups of words in the same ‘conceptual community’.

Step 4: Analysis of topics

Last, we analyzed and ‘annotated’ by hand the topics produced by blind LDA and its semantically (i.e. ontology) enhanced version through LogicLDA, respectively. TM performed via LDA (both in its blind form or in the semantically enhanced one) identifies topics that are simply a collection of words. The task of linking such lists of words to a specific theme is left to the user/interpreter. Our work aims precisely at facilitating this task, by devising a method that produces topics already aligned with the knowledge encoded in an ontology. Our analysis verified that the topics retrieved via the semantically enhanced LDA were easier to categorise.

Results and discussion

Figure 2 shows a summary of the topics extracted with blind LDA (left) and its ontology semantically enhanced version (right). In both cases, we chose to detect 33 topics, corresponding to the number of tightly linked clusters in the ontology. The first result to highlight concerns the overall structure of the topic map. Indeed, the inclusion of the ontology substantially changes the structure of the topics retrieved. In Figure 3 we show a matrix reproducing the pattern of word pairs co-occurrence within topics: each row/column of the matrix is a word, darker elements of the matrix imply a more frequent co-occurrence in the topics of the row/column word pair. The matrix on the left shows the results of standard LDA, while the matrix on the right those of the ontology enhanced one. The order of words in the two matrices is preserved. As it can be observed, the co-occurrence patterns visibly vary between the two cases, with the ontology-enhanced LDA showing higher word co-occurrence.

The second result to stress is connected to the content of single topics. When the number of topics exceeds the heterogeneity of the corpus, LDA has the shortcoming of yielding some transversal topics that gather all texts which are difficult to classify. These transversal topics do not focus on some specific theme and are hard to interpret. We show in Figure 2 those transversal as uncoloured circles. The ontology enforces a given scientific perspective and reduces the number of ‘transversal’ topics that are retrieved by LDA. Since the ontology forces LDA to find topics related with the themes treated in the ontology, the number of transversal topics decreases from 5 to 3 when using the ontology.

The third notable result has to do with the focus of the topics: the introduction of the concept dealt with by the ontology forces some macro-themes uncovered by blind LDA to become much more defined. This is the case, for instance, of the topic dealing with policies, that becomes specifically an education policy topic when introducing the ontology.
Figure 2 The topics extracted with blind LDA (left) and the semantically enhanced version of LDA (right). For each topic, we show the first five characteristic words. The size of the dots is proportional to the topic incidence in the corpus, transversal topics have no filling.

Figure 3 Matrix of word cooccurrence in topics, for the blind LDA (left) and semantically enhanced LDA (right). The matrix shows co-occurrence for all pairs of words: darker tones of blue imply a more frequent occurrence of the word pair across topics. Since the sequence of words in the matrices is the same, comparing the two matrices allows to appreciate the difference between the overall structure of the topics in the blind LDA case and in the ontology enhanced one.

Conclusions
We presented in this paper a first attempt at integrating a domain ontology into LDA topic modelling. We did so by focusing on the issue of “obesity”, by building our own ontology on the theme and by leveraging on LogicLDA to enforce the logical constraints encoded in the ontology.

Our preliminary results suggest that incorporating domain knowledge into topic modelling is indeed viable and that the results obtained via standard LDA and its ontology informed variant do differ. Furthermore, we found that the classification of texts obtained with an ontology is
easier to understand (since they are classified according to scientific perspectives) and that the selection of the logical clusters in the ontology enhanced the retrieval of some specific subtopics of interest (since the ontology encompasses a plurality of views).

In the light of these promising preliminary results, our goal is now to fine tune and systematise the method devised, in order to offer a tool to categorise scientific research in a way that is easier to interpret according to the specific societal needs represented by the ontology.

References


Bibliometric study on big data research: An integration of topic model and citation network analysis

Dong Ke¹, Wu Jiang², Cheng Ni³,

¹dongke@whu.edu.cn
School of Information Management, Wuhan University, Wuhan(China)

²jiangw@whu.edu.cn
School of Information Management, Wuhan University, Wuhan(China)

³ncheng@mail.hzau.edu.cn
Huazhong Agricultural University, Wuhan(China)

Abstract
Big data has been attracting wide attention due to its great importance. The rapid development of big data in recent years has led to a large amount of publications containing the achieved knowledge of this area. To study the intellectual structure of the related research about big data, a retrospective bibliometric analysis is conducted based on the Web of Science databases. 13673 papers and 8016 citation links are collected. The LDA topic model are used to detect the topic distribution of big data research area. 11 topics about big data technology, application and security are found. Island algorithm is applied to find the most influential 28 research communities from citation network, and these communicates are labelled through topic distribution rather than traditional way of single tag. Labelling each cluster by topic distribution can reveal the research content for sub-structures, which reflects a much more comprehensive picture of big data research area and provide a valuable reference for researchers to understand the overview and present situations in this field.

Keywords
Citation and co-citation analysis; Topic model; Social network analysis; Cluster; Big data; Algorithm

Conference Topic
Citation and co-citation analysis; Social network analysis; Mapping and visualization;

Introduction
Structured and unstructured data is produced massively in scale under the constant development of information technology and internet environment. Relevant statistics show that a social networking site, Twitter, has a record 500 million pieces of daily new Tweets (Singh et al., 2015). Those data, large scaled in various types and updated very frequently, can no longer be explained by traditional definition of data, which results in the birth of ‘big data’ at the right moment. As early as 2001, Laney, an analyst at Gartner,
a US information technology consultancy, analyzed the characteristics of massive
growth data from three perspectives: volume, velocity and variety (Laney, 2012). Since
then, big data’s value density (LaValle et al., 2011), authenticity (Raghupathi et al.,
2014), complexity (Fan et al., 2013) and other features are further explained, and the
value of large data has been constantly recognized. The United Nations Global Pulse,
in a 2011 report, points out that big data has greatly improved the efficiency of decision-
making in key areas of social development such as health care, employment, and
economics, but the challenges that it brought with are not to be overlooked (UN Global
Pulse, 2012). At present, the government, academia and the business community
maintain an intensively increased attention to big data. Big data has become a new
research area which develops rapidly and produces results fruitfully.

In this study, a bibliometric analysis on big data research area is to be made based on
big data research papers retrieved from Web of Science database without timespan
settings which enables to explore initiating papers with big data concept. The analysis
methods mainly include LDA topic model and citation network analysis. LDA topic
model is used to extract the topic distribution of the whole research area; meanwhile,
citation network analysis is used to detect highly attractive communities of big data
research. This study also proposes a new method to integrate the topic model analysis
results and structural analysis results of citation networks. This method can effectively
reveal the communities which attract widely concern.

This paper is organized as follows: the literature review part introduces big data
research and related methods, such as citation network analysis and content based
citation analysis; Methodology explains data source, retrieval strategy, LDA topic
model, communities detection algorithm and a proposed method that integrates topic
model and citation network analysis; Results shows the topic distribution of big data
research area, as well as the topic distribution of highly impact communities which are
detected. Conclusion sum up the whole study and discussion presents views to future
related research.

Literature Review

Brandl, a scholar of University of Vienna, in his early paper on big data published in
2004, proposed to use knowledge mining to find economic data and forecast the trend
of macroeconomics (Brandl B., 2004). In September 2008, Nature published a special
research theme on big data, and a number of articles discussed the influence of big data
on modern life and scientific research (Lynch, 2008). In 2012, many international
organizations and government departments issued initiatives and reports related to big
data, such as ‘big data, big impact’ report released the Davos Economic Forum (World
Economic Forum, 2012) and ‘Big Data Research and Development Initiative’ issued by
The Obama Administration (Executive Office of the President, 2012). These initiatives
and reports have aroused widespread concern in society and have produced a great deal
of research findings. Singh et al. (2015) once used scientometric method to analyze big
data studies in terms of major countries, institutions, authors, grow of output, authorship
and country-level collaboration patterns and top publication sources. They used the
distribution of papers’ classification and some selected control terms to analyze the
subjects of big data research area. Singh’s study retrieved and analyzed only 1416 papers in WOS and 6810 papers in Scopus from 2010 to 2014. But papers related to big data have rapidly increased to more than 10,000 by the year of 2016. Fig. 1 shows the time distribution of published research papers on big data. It is found that research papers on big data begin to grow rapidly since year 2012. Thus, this paper aims to update the previous research study and to effectively reveal the current situation of big data research area.

Citation networks are sequentially dependent: more recent publications can only cite older ones. Citing behavior among scientific publications can reveal dynamic knowledge flows. So the complex and inter-connected structure of the citation networks has been utilized for studying the knowledge structure of a given research field. Using citation based approaches to analyze the intellectual structure of science has a long history. Some outstanding former studies used clusters of highly-cited papers to represent the scientific specialties and explored the entire specialty structure of science (Small and Upham, 2009) and research fronts (Aris et al. 2009). Other approaches include co-citation analysis among highly cited authors, journals, or keywords. Mapping all the scientific literature, sometimes so called knowledge domain visualization has gained increased attention in recent years (Chen 2003). Meanwhile, with the fast development of social network analysis, more and more approaches for whole network analysis have been employed (Snir & Ravid 2016; Tang & Tsai 2016; Tang et al. 2016). Since Ding et al.(2014) introduced the content based citation analysis, text mining methods has been widely used in citation network analysis (Chen 2017; Jha et al. 2017; Kim et al. 2016).

**Methodology**

*Data pre-processing*

The data for the present study were acquired from the on-line edition of the science citation index-expanded (SCI-E), social science citation index (SSCI), and conference
proceedings citation index (CPCI) database published by Thomson Reuters. A set of publications (articles, reviews, proceedings papers and software reviews) related to big data were generated by searching the term ‘bigdata’ or ‘big data’ in the bibliographic field “TS” of the ISI Web of Science database. Papers about large scale data array and large scale data analysis were also included, but the meaning of big data mentioned in these papers was far way different from those in current researches. After repetitious papers eliminated, 13673 papers were collected from 4958 journals, proceedings, and books. The data were retrieved on March 25, 2017.

To detect the topics of big data research area, title, abstract, keywords and keywords plus are used. Citations among the 13673 papers are also cleaned for analysis. Citation network is acyclic because a paper can only cite early published papers. As a result, all arcs in the citation network point backwards in time. However, it was found that there were two exceptions: some papers published earlier cite papers published later; also, some papers are of self-cited. The reasons for the former exception have been discussed by Rousseau and Small (2005), who gave an extreme case which is called Escher Staircase phenomenon. The later exception is caused by the mistakes from Web of Science or by authors who have made papers self-cited. In this study, citations of the former exception were re-arranged so that they are chronically in order according to the publication time; citations of the later exception were removed. Finally, 8016 citations were collected for the generation of citation network.

Research methods

**LDA topic model.** The topic extraction of the text content is to select the appropriate text content topics and characteristic vocabulary in order to describe characteristics of the text content. Topic model, a probability generation or production model of text content, such as latent semantic analysis (LSA) (Deerwester et al., 1990), probabilistic latent semantic analysis (PLSA) (Hofmann, 1999) and LDA (Blei et al., 2003), which can express the meaning of the text to an extreme and solve semantic relevance problems among vocabularies, topics and text by simulating human thinking process, is the most commonly used text topic extraction method. LDA topic model adopts Dirichlet distribution, simplifies the derivation of the model and avoids the over-fitting problem of the LSA and PLSA models (Hofmann, T., 2001). Therefore, it has a good priori probability hypothesis that the number of parameters does not grow linearly with the increase of the number of texts with a strong generalization ability and excellent performance in algorithm complexity algorithm complexity and exhibition effect, the derivative model is widely used in the fields of topic mining, classification clustering, retrieval and topic evolution of text analysis (Yang et al., 2016; Tan et al., 2016; Morchid et al., 2017). This study uses the Incremental Hierarchical Dirichlet Processes to realize topic detection (Gao et al., 2011).

**Highly impact communities detection from citation network.** Island algorithm is applied to detect communities (clusters) from citation network in this study. According to Nooy et al. (2011), an island cluster is a maximal sub-network of vertices connected directly or indirectly by lines with the value of vertices in the cluster greater than those outside
the cluster. Island algorithm has its advantage in mining the cluster in network without basing on one single standard. It concerns not only the size of sub-network, but also vertices of greater importance than those around communities. This is helpful when to detect topics that are not so hot but has formed to a large scale, especially when to analyze exceptionally large-scale connected component.

For example, in a citation network composed by research papers on information retrieval, these papers form a large connected component which are about algorithms and user behavior. In this largely connected component, all papers about algorithms are highly cited (e.g. all of their citation scores are larger than 100), whereas those on user behavior are lower cited (e.g. the largest citation score is 50). To detect communities in such a big component, if a standard threshold of citation count is set at 70, community of user behavior will be ignored; whereas, if the standard threshold is lowered at 30, though the community about user behavior shows up, the community of algorithms will be severely expanded to such an extent that many non-essential research papers will be flooded into the network. Fortunately, island algorithm can solve this problem. It can hold many highly impact communities at the same time, and each of them has similar size and is formed by research papers with higher citation scores than those papers who around the communities. In this paper, vertex island algorithm will be used (Nooy et al., 2011).

**Topic distribution of Clusters.** The LDA topic model can be used to find the distribution of words in topics and the distribution of topics in documents, that is, it can extract topic from words and documents respectively. In a literature cluster, the weight of each document is different (e.g. they have different betweenness centrality), so the weights of topics in each cluster cannot simply be the sum of weights of each topic, but need to be converted at first. Suppose that cluster C contains n documents, i.e. C=(d₁, d₂, ..., dᵢ), and these documents are represented by m topics, let the weight of the document dᵢ is wᵢ, and the weight of dᵢ for topic k is pᵢk. Let Xₖᵢ as weight for topic k after the document dᵢ is converted, then we get

\[
\begin{bmatrix}
X_{11} & X_{12} & \cdots & X_{1m} \\
X_{21} & X_{22} & \cdots & X_{2m} \\
\vdots & \vdots & \cdots & \vdots \\
X_{n1} & X_{n2} & \cdots & X_{nm}
\end{bmatrix}
\begin{bmatrix}
w_1 \\
w_2 \\
\vdots \\
w_n
\end{bmatrix}
= 
\begin{bmatrix}
p_{11} & p_{12} & \cdots & p_{1m} \\
p_{21} & p_{22} & \cdots & p_{2m} \\
\vdots & \vdots & \cdots & \vdots \\
p_{n1} & p_{n2} & \cdots & p_{nm}
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{bmatrix}
\]

Besides, the calculation of the weight Yᵢ of cluster C in topic i is as the following:

\[
Y_i = \sum_{k=1}^{n} X_{ik}
\]

Traditionally, analyst always gave each cluster only one label of its content. For example, in Citespace, 4 algorithms such as LIS, TF*IDF, LLR and MI are used for labelling one tag (Chen 2017). Actually, one cluster is a sub-field of a research area. One tag labelling can help to filter out unnecessary information and find out main subject of a cluster. But one cluster, i.e. one sub-field always cover more than one subject. This study tries to use more tags to show the topic distribution of every cluster, which can attribute to a comprehensive analysis.
Result.

**Topic distribution**

11 topics were selected finally after a number of trials. As higher-frequency keywords in each topic were given in the result of LDA model analysis, the topics can be named according to these keywords as it is shown in Table.1. In general, current research on big data mainly focuses on three categories- core technology of big data, application of big data in many areas, and personal privacy protection in big data environment. In the category of core technology, the main research topics at present include Distributed System, Algorithms, Database, Semantic web and knowledge discovery, and Machine learning. In the category of big data application, the related topics include Smart city, Medical science, Social media and mobility network, Remote sensing and other applications area such as business, education and manufacturing. Scholars mostly concerned about the security issues in the big data environment focus on the protection of personal privacy.

<table>
<thead>
<tr>
<th>Category</th>
<th>No.</th>
<th>Topic</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>1</td>
<td>Distributed System</td>
<td>mapreduce, hadoop, distributed, storage, parallel, applications, memory, time, algorithm, scheduling</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Algorithms</td>
<td>algorithm, dimensional, cluster, proposed, problem, methods, approach, large, optimization, datasets</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Database</td>
<td>analytics, database, applications, framework, storage, distributed, nosql, software, architecture</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Semantic web and knowledge discovery</td>
<td>mining, knowledge, semantic, web, search, text, graph, method, process, approach</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Machine learning</td>
<td>learning, machine, classification, image, prediction, feature, detection, neural, proposed, recognition</td>
</tr>
<tr>
<td>Application</td>
<td>6</td>
<td>Smart city</td>
<td>network, smart, traffic, energy, power, networks, monitoring, sensor, grid, city</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Medical science</td>
<td>healthcare, clinical, medicine, patients, disease, treatment, cancer, drug, biological, biomedical</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Social media and mobility network</td>
<td>social, media, network, online, twitter, public, mobile, user, sentiment, behavior</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Remote sensing</td>
<td>spatial, sensing, remote, image, land, climate, earth, images, GIS, geospatial</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Other application areas</td>
<td>management, education, industry, service, application, internet, business, innovation</td>
</tr>
<tr>
<td>Security</td>
<td>11</td>
<td>Privacy security</td>
<td>security, privacy, internet, service, mobile, services, IOT, devices, access, users, encryption</td>
</tr>
</tbody>
</table>
The distribution of each topic in the dataset is shown in Figure 2. The length of the bar represents the cumulative frequency of the keywords in the topics. It can be found in Figure 2 that the topic related to technology accounts for the majority of the current research on big data, and the topics of Distributed System, Algorithms, and Database are the largest in numbers.

Figure 2. The distribution of each topic in the dataset

To the topics related to technologies: (1) Distributed system. MapReduce is an important program for big data concurrent applications and is a core component of Hadoop. Distributed computing programming can be implemented rapidly on Hadoop platform through MapReduce, so the research in the topic of Distributed System is closely related to MapReduce and Hadoop. (2) Algorithms. As there are more information dimension in big data, a large number of papers focused on high-dimensional data processing in the topic on Algorithms. For example, there are many papers on implementation of efficient cluster in big data. (3) Database. The traditional database cannot meet the demand of storage and processing of large scale data in big data environment. Therefore, big data solutions based on NewSQL and NoSQL database received wide attention and scholars studied on big data storage and analysis through new database platforms. (4) Semantic web and knowledge discovery. In traditional information environment, the level of collaboration between Semantic web, Linked data communities and the Applied Ontology community has been much less than expected whereas big data can well connected these communities and provide opportunity for the development of Semantic web. So the topic related to Semantic web becomes an important domain among current big data research. (5) Machine learning. The core of big data is to use the value of data. Machine learning is key technology of using data value and is thus indispensable for big data. For machine learning, the more the data is, the more chances are to improve the accuracy of the model. Big data provides new experimental environment for machine learning. At the same time, machine learning, as important data analysis technology, has attracted attention by
To the topics related to application: (1) Smart city. The infrastructure including smart grid, urban energy management, real-time monitoring and forecasting of urban traffic flow needs the strong support from big data. Big data application in Smart city is most prominent. (2) Medical science. There are many fruits of big data application in the field of Medical science. At present related big data case sets on diseases have been established, which can support dealing with emergency situations and improvement of nursing quality during the patient care process. (3) Social media and mobility network. A large amount of data produced during the wide application of social media meets the core features of big data such as large volume, rich variety, low value density, and fast processing speed. As digital footprint in human life and communication, social media and mobility data have become an important kind of big data. Social media data can provide data basis for research on human behavior, policy influence and economic development. (4) Remote sensing. The data produced in remote sensing activities is rich and typical characteristics of big data. Big data technology provides important solutions for complex remote sensing data processing in the fields of weather forecasting, water resources monitoring, and environmental protection. (5) Other application areas. Big data can be also used in business and education. Big data and business intelligence is paid more attention to improve strategic decision making efficiency in the business field. How to use big data to improve higher education quality and student information literacy is hot spot in the field of education. 

To the topic about security issues. The information and privacy risks generated with the continuous development of big data are rising. Research on risk control is significant. At present, privacy protection in the big data environment is in initial stage. The related research focused on specific technical means and tools.

**Communities in Citation network**

![Figure 3. Mapping the clusters in citation network](image)

Here we use the betweenness centrality scores for island cluster detection. The betweenness centrality of a node in the network measures the importance of the position...
of the node in the network. Those paper that highly connected to other nodes such as hubs or positioned between different groups of nodes may have high betweenness centrality scores (Chen 2017).

Figure. 3 is the mapping of the whole network and island clusters. 8635 papers are not cited by other papers, in other words, only 5058 pieces of papers are paid by researchers’ attention and cited. There are 300 connected components with size >2, and the largest component is constructed by 4334 papers. 28 clusters are extracted by using island algorithm as it is shown in the right part of Figure 3. Size of each cluster is set in a range from 5 to 15. These 28 extracted island clusters are 28 communities that attract more attention than other communities. Labeling these clusters is next step for analysis.

**Figure 4. Topic distribution of 28 communities**

The topic distribution of communities as it is shown in Figure 4. All the communities can be classified into 3 types. All the papers of the first type are focused on one single topic. For example, community 17 in Figure 4 which consists of 12 papers only focused on medical science. Papers belonged to this cluster discussed how to use big data to improve Emergency Care, cardiovascular care, in the Care of Children, big data application in European health systems and databases are diverse and fragmented. Technologies are scarcely mentioned or discussed.

To the second type, such as community 1 in Figure 4, mainly focus on a few topics. Though the weight of the topic “medical science” is the largest, unlike cluster 17, papers of this cluster are more concerned on how to obtain information from Internet and social media to improve medical research and healthcare. For example, MacFadden et al. (2016) developed a Web-based/mobile platform for aggregating, analyzing, and disseminating regional antimicrobial resistance information. Lee (2016) Harnessing Spatial Big Data for Infectious Disease Surveillance and Inference.

To the third type, such as community 27, the research contents nearly relate to all topics. Community 27 includes 11 papers which have high scores to all the topics. It is found that most of them are researches at a macro level which mainly focus on the technological changes by big data environment and its future, framework of
This study explores the intellectual structure of the big data research through examining papers indexed by Web of Science. 13673 papers and 8016 citation links among them are collected to build a citation network. This study use LDA topic model to analyze the topic distribution of the dataset, and classify different topics into three categories: core technology, application and security issues. The core technology in big data research mainly focus on Distributed system, Semantic web and knowledge discovery, Machine learning, Algorithm and Database. Application areas include Smart city, Social media and mobility network, Remote sensing, Medical Science and other application areas such as business, education, and so on. Besides, security issue is mainly about privacy security.

Through island algorithm, 28 island clusters are detected from the citation network. These clusters are communities who attracted higher attention in some sub-fields and form connected components at large scales. Then by using topic model, topic distribution of each cluster is attained. Though some cluster has papers on single one topic, most clusters cover more than one topics. Compared with the traditional way of using one tag for cluster labelling, more tags result in showing a comprehensive picture as well as the diversity feature of big data research.

**Discussion**

LDA topic model can detect the topic distribution of a research area, but how these detected topics are linked to each other? To tackle this question, citation network analysis can be adopted to integrate with LDA topic model. Citation network reflects real connection in the knowledge diffusion process. If to put focus on topic co-occurrence in island clusters, a much more specific and real connection among topics will be reflected. Take researches on technological application of a field for example,
if all the topics are classified into two categories: one is technology and the other is application (privacy security is also included), topic co-occurrence will be detected, as it is shown in Figure 5. It is found that those technologies are widely used in different areas.

![Figure 5. the bipartite graph of topic co-occurrence](image)

LDA topic model which is able to detect the distribution of dataset, but it fails to determine the existence of real relationship among papers. The same word or topic can only mean the similarity may exist between different papers, but can never testify knowledge exchanges happen in real world. Citation reflects real research communication but analysis based on citation network does not take research content into account. Thus, if to combine LDA topic model and citation network analysis, a better analysis result can be yielded. However, interpreting the results requires circumspectness due to some limitations of this study. In addition to datasets from Web of Science, open access journals and online publications are also can be considered to reflect the true structure of big data research field better. Also, the number of topics and island clusters depends on pre-set parameters. Proper number of topics and island clusters needs to be determined by repetitious trials, which will influence the research efficiency.

**Acknowledgments**

This paper is supported by the National Science Foundation of China (Grant No. 71603195), China Postdoctoral Science Foundation (Grant No. 2016T90735) and National Social Science Foundation of China (Grant No. 15CTQ024)

**References**


World Economic Forum. (2012). *Big data, big impact: new possibilities for international development.* from


New Citation Analysis Perspective Based on Ontology and Linked Data

Shi Zeshun\textsuperscript{1}  Xiao Ming\textsuperscript{2}

\textsuperscript{1}shizeshun@gmail.com  \textsuperscript{2}ming_xiao@bnu.edu.cn

Beijing Normal University, School of Government, 100875, Beijing (China)

Introduction

Citation analysis is a bibliometric analysis method which reveals the quantitative characteristics and laws of scholarly publications. It involves the use of mathematical and statistical methods to analyse citations within journals, papers, authors, and other references. Citation analysis has seen substantial theoretical and practical progress over several decades of development, and has been widely applied to evaluate scientific knowledge, identify scientific models and detect new frontiers which being explored by the scientific community.

Traditional citation analysis methods and tools are overly dependent on citation databases, which have the following drawbacks:
1. All citation acts are treated as equally important;
2. All kinds of statistical indicators are based on specific instances of reference, which are annotated only by the author;
3. Citation databases can only be used to reveal citation relation between different papers, but fail to reflect any deeper relationships among them.

Motivations and behaviours related to citation have been analysed by researchers from various angles. Content-based citation analysis method has also been proposed by Ding et al. (2014). In this paper, we propose a new citation analysis framework based on ontology and linked data; our goal is to enhance the efficacy of citation analysis via semantic web technology.

Related work

In recent years, researchers have begun to introduce semantic web technology to citation analysis in effort to exploit ontology, linked data, and other technologies to improve the description of citation behaviours and motivations. The most representative example is the Semantic Publishing and Referencing (SPAR) Ontologies created by Shotton, Portwin, Klyne, and Miles (2009). CiTO ( Citation Typing Ontology) is the ontology which SPAR uses to describe the relationship between citing papers and cited papers; it provides reference information such as background, method, citation type (e.g., journals, books, reports), peer review, and so on.

Other researchers, for example, Ding, Konidena, Sun, and Chen (2009), have explored the idea of semantic citation to suggest that individuals can use ontology and linked data to describe bibliographic data and publish it to RDF triples. Mahmood, Qadir, and Afzal (2013) combined semantic web technology with credible citation analysis to establish a framework that provides openness and reliability validation for all stages of the citation behaviour lifecycle. The framework makes author, publisher, database vendor work together, thus building a set of reliable reference information while eliminating any false or misleading citation actions in the literature.

Method and process

Any citation analysis method based on ontology and linked data mainly includes the following three steps: First, building citation ontology. Then, using the citation ontology to normalize the reference information and publishing the citation data to citation linked data. Finally, writing specific SPARQL search queries to extract the data needed for citation analysis.

Citation ontology construction

From the perspective of citation analysis, bibliographic citation information and full-text citation information are two independent parts but also two important sources of data that are both necessary for citation analysis. So, we construct the Bibliographic Citation Ontology (BCO) and Full-text Citation Ontology (FCO) based on the bibliographic citation information and the full-text citation information, respectively. This allows us to achieve comprehensive semantic annotation of the citation information at hand.

From the list of references, information such as the author, periodical, document type, year, volume period, and page number are extracted as the classes and properties of BCO. On the other hand, the construction of FCO begins with three aspects: citation function, citation sentiment and citation position. Citation function represents the role of cited work to citing work, such as background development, data support, methodology support, extension, refutation and so on. Citation sentiment express the emotion attitude from citing work to cited work, such as positive, neutral, and negative. Citation position indicates the location of the paragraph where the reference behaviour occurs, such as the "Introduction" or "Conclusion" part of the paper.

An example of the FCO ontology’s classes and properties is shown in Figure 1.
Publishing citation linked data

Then, we can convert citation information to citation linked data according to citation ontology. We recommend D2R as the linked data conversion tool for this purpose. D2R is a very popular tool for linked data publication which serves to convert the massive, relational database format data into linked data RDF triples. After that, we need a database to store citation linked data. Both OpenLink Virtuoso and Apache Jena Fuseki are very good choices.

Citation analysis method implementation

The essence of the new citation analysis is to write the corresponding SPARQL queries, which can be used to extract the citation information of specific dimensions. The search results are then calculated and visualized to analyse the citations per different dimension. For example, co-citation analysis, bibliographic coupling analysis, citation language analysis, citation country analysis, citation age analysis, citation journal analysis, and so on are based on the bibliographic data of traditional citation analysis, while citation function analysis, citation sentiment analysis and citation position analysis are based on a full-text citation analysis perspective.

We write SPARQL queries to extract specific citation information, which is shown in Figure 2. The first SPARQL query is used to retrieve all the publication year information for the references cited by paper A, and the second query to retrieve all references to reference 4, which extends the function of document 4. The search results are then calculated and displayed as the final results. Visualization software (e.g., Graphviz, Relfinder) could also be applied to simplify the display of results. Other dimensions of citation analysis can be implemented according to the same principle. As the quality of data is continually improved, more dimensions of citation analysis can also be achieved in follow-up experiments.

Conclusion

In this paper, we proposed a new citation analysis framework based on ontology and linked data. By combining these technologies into a new semantic web with citation analysis method, we hope to improve the traditional citation analysis method (which relies heavily on citation databases). With the rapid development of semantic publishing and projects like the Open Citation Corpus (OCC), it is possible to mark massive amounts of citation information as machine-readable RDF triples. In the future, we plan to design further experiments to verify the feasibility of the proposed method. We hope that introducing ontology and linked data into citation analysis will yield optimal results while facilitating new technological development and innovations.

Acknowledgments

This work was supported by a grant from the National Social Science Foundation of China (No. 16BTQ073).

References


Introduction

Many public departments produce or collect a broad range of different types of data during the process of performing their tasks (Zhou & Ma, 2016). The large quantity and centrality of data collected by governments make these data particularly significant as an important information resource for the public (Cai & Xia, 2015). Since 2009, lots of movements have pushed governments to publish relating policies and principles to guarantee the constant development of open government data (OGD) (Ubaldi, 2013). In 2015, State Council of the P.R. China has published a policy for the promotion of the development of big data, which became the guideline for OGD development in China. Under its influence, a lot of local OGD platforms are built. How to get a thorough understanding of the present development of OGD in China has become a focus. Though researchers in China have made efforts to evaluate OGD websites, little attention has been paid to the context and impact of OGD. Thus, this study aims to build a reasonable model that fits the assessment of OGD development in local area level. So that we could get a thorough view of the strength and weakness in the present development of OGD in different areas in China.

Method

The evaluation framework

In order to build such an evaluation framework, we first investigated into the evaluation systems worldwide for their structures and variable selections. We selected 8 evaluation systems as samples: Open Data Census, Open Data Index, Open Data Barometer, Open Government Index, Open Data Readiness Assessment, OUR Data Index, Open Government Budgetary Data in Brazil, Open Data Monitor. These systems are: Firstly, complete and systematic, having a frame for the OGD evaluation. Secondly, they have been applied to OGD evaluations for many years, thus they are practical. Based on the investigation of eight OGD evaluation systems, we proposed a three-layer framework as the basic of the assessing model which was shown in Table 1. The first layer listed three main aspects of the evaluation, while the second and third layers contain specific variables. Readiness focused on the conditions for OGD development from the view of local government. Dataset focused on the nature and qualities of open datasets, including the openness of data, issues of data quality and quantity. Impact focused on the benefits to be gained from the OGD.

<table>
<thead>
<tr>
<th>Table 1. The evaluation framework</th>
</tr>
</thead>
</table>

Analytic Hierarchy Process

The Analytic Hierarchy Process (AHP) is a general theory of measurement through pairwise comparisons and relies on the judgements of experts to derive priority scales (Saaty, 2008). It was developed by T.L. Saaty in 1971-1975. The study used AHP to get the priorities of all the elements in the above evaluation framework, and used Yet Another AHP Software (Yaahp 10.5) as an important aid to AHP analyses.

Procedure

First, we transferred the evaluation framework into the hierarchic structure in AHP, as was shown in Figure 1. Secondly, we created the expert questionnaire for the pairwise comparisons of the importance of elements that belong to each of the upper level elements in the hierarchy. We chose 10 experts in the area of OGD research in China to fill in the questionnaires. They ranked the importance of each pair elements from 1 to 9. After we collected expert questionnaires, we first examined the consistency of each of the comparison matrices. We abandoned those matrices with the consistency ratio (C.R.) over 0.10, and then calculated the weighted geometric mean of the values in each matrices given by different experts to form the final matrices for

1600
calculating the priorities. After that, we derived the local scale of priorities by solving for the principal eigenvector of the matrix and then normalizing the result. We continued this process of weighing until the final priorities of the alternatives in the bottom most level are obtained. After weighting by the priority of its parent criterion (which for the second-level elements is always equal to unity, the weight of the focus), we finally got the global derived scale for each variables in the model. Yaahp 10.5 was used to make these calculations and guided us to improve the inconsistency if needed.

Figure 1. Hierarchy with weights

Results

Following the above procedure, we first got the final matrices after calculating the weighted geometric mean of values in the same pairwise comparison matrix graded by different experts. We also examined the C.R. in each matrix to make sure of its consistency. Then we derived the local scale of priorities (W_i). After that, we weighted the local scale by the priority of its parent criterion, and got the global derived scale of each of the elements (Marked in Figure 1).

An Initial Evaluation

In order to test and verify the performance of the evaluation model, we selected Shanghai’s OGD platform to carry out the initial evaluation. Table 2 listed the value of each elements in the evaluation model as well as its weight derived from the AHP process. We could find that Shanghai got a good performance in their data access and data quality. But the government in Shanghai need to pay more attention to the support of business utilization of OGD, as well as the protection of personal data. The evaluation also shows that the OGD did not have a great impact on the economy development as well as improving government efficiency.

Discussion

The assessing model has an advantage of analysing the strength and weakness of the OGD development in China. The model could also become a good aid for setting goals or making plans before and during the development. Another advantage of this evaluation model is that it covers the whole developing process, from preparing the overall environment to the impact of OGD on different aspects. Thus, it guaranteed a relatively long period of validity for this model.

We also noticed that the quantity variables have a great influence on the final accessing grade because their values are too large compared with other variables. Thus, in future evaluation process, it is better to find a possible solution to reduce the numerical disparity between quantity variables and quality variables.

Table 2. Evaluation Result of Shanghai

<table>
<thead>
<tr>
<th>Variables</th>
<th>Value</th>
<th>W</th>
<th>Result</th>
<th>Variables</th>
<th>Value</th>
<th>Aspects</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>B11</td>
<td>78.5</td>
<td>0.0528</td>
<td>4.1448</td>
<td>B11</td>
<td>10.3848</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B12</td>
<td>78</td>
<td>0.0809</td>
<td>6.2409</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B121</td>
<td>80</td>
<td>0.0963</td>
<td>0.5580</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B122</td>
<td>62</td>
<td>0.3669</td>
<td>2.5206</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B131</td>
<td>77</td>
<td>0.0187</td>
<td>1.4999</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B132</td>
<td>54</td>
<td>0.0073</td>
<td>0.49254</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B133</td>
<td>46</td>
<td>0.0010</td>
<td>0.0866</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B211</td>
<td>579.7159</td>
<td>0.0169</td>
<td>9.7972</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B212</td>
<td>918</td>
<td>0.0254</td>
<td>24.4812</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B213</td>
<td>31</td>
<td>0.0186</td>
<td>0.4272</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B214</td>
<td>100</td>
<td>0.0286</td>
<td>2.6800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B221</td>
<td>100</td>
<td>0.0621</td>
<td>6.2100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B223</td>
<td>100</td>
<td>0.0375</td>
<td>3.7500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B224</td>
<td>100</td>
<td>0.0399</td>
<td>1.9900</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B233</td>
<td>100</td>
<td>0.0381</td>
<td>3.8100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B234</td>
<td>100</td>
<td>0.0608</td>
<td>6.6800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B235</td>
<td>100</td>
<td>0.0343</td>
<td>3.4300</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B236</td>
<td>100</td>
<td>0.0355</td>
<td>3.9042</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B237</td>
<td>100</td>
<td>0.0466</td>
<td>4.6000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B238</td>
<td>100</td>
<td>0.0269</td>
<td>2.6900</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B239</td>
<td>75</td>
<td>0.0183</td>
<td>1.3575</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B241</td>
<td>100</td>
<td>0.0236</td>
<td>2.3600</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B311</td>
<td>88</td>
<td>0.0093</td>
<td>0.0324</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B312</td>
<td>86.5</td>
<td>0.0227</td>
<td>1.9656</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B313</td>
<td>81.5</td>
<td>0.0306</td>
<td>2.4939</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B314</td>
<td>81.5</td>
<td>0.0651</td>
<td>5.3057</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B315</td>
<td>77</td>
<td>0.0451</td>
<td>3.4727</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B316</td>
<td>72.5</td>
<td>0.0732</td>
<td>3.2070</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B317</td>
<td>76</td>
<td>0.2073</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B318</td>
<td>74</td>
<td>16.4253</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B319</td>
<td>78</td>
<td>19.4017</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B320</td>
<td>72</td>
<td>19.7152</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reference


BIBLIOMETRIC ANALYSIS AND RANKING OF LIBRARY AND INFORMATION SCIENCE (LIS) RESEARCH AND PUBLICATIONS IN AFRICA

Okon E. Ani1  Eucharia Okwueze2

1 anioedet@yahoo.com
University Library, University of Calabar, Calabar, Nigeria

2 okwuezeeucharia@gmail.com
Department of Library and Information Science, University of Calabar, Calabar, Nigeria.

Introduction
Bibliometric analysis is a method that is used to describe patterns of publication and national and international strengths and biases in areas of research within a given field or body of literature such as library and information science (LIS) (Aharony, 2011). Johnson (2011) explained that many bibliometric studies enumerate the extent and scope of scientific communication amongst scholars/researchers. According to Ocholla and Ocholla (2007), bibliometric studies are widely used to inform policies and decisions in political, economic, social and technological domains affecting information flow and use pattern within, between and outside institutions and countries. Thus, bibliometric studies are essentially used in evaluation of research and publications in all areas of human endeavors to support decision making by policy makers in governments, organizations/institutions (universities), academic departments and individual scholars/researchers. Furthermore, bibliometric analysis is significantly used in the ranking of research and publications (Meho and Spurgin, 2005) and this has become popular especially in the universities and other academic and research institutes with emergence of different rankings organizations within the past one or two decades (Ani, 2015). Barik and Jena (2016) opined that bibliometric study is focus on evaluation of research performance of authors, institutions/universities, countries or regions in variety of disciplines. Publications counts and citation analysis are two major methods that are used in bibliometric studies. Ocholla and Ocholla (2007) described citation analysis as a tool that is used as quantitative measure of the quality of research and publications.

LIS is a discipline that is fast growing in terms of research and publication globally. Although bibliometric analysis and ranking studies in LIS have a long history in developed countries (Meho & Spurgin, 2005); in Africa, bibliometric studies are limited and insignificant in Africa in spite of the widely applications of these studies by policy makers to support research and publications in line with the emerging global rankings of universities (Ocholla & Ocholla, 2007). Onyancha (2007) vividly captured the poor state of LIS research in Africa by expressing the fear that there is general concern among LIS scholars concerning the growth and development of LIS research in Africa in relation to global perspective. Thus, this present study is intended to fill the knowledge gap in bibliometric studies and would therefore expand the scope of literature in LIS in Africa.

Objectives of the Study
The study was guided by the following objectives:
1. To rank LIS research and publications in African countries in global perspective;
2. To rank LIS research and publications in Africa by country, and
3. To determine trends in LIS research and publications in Africa.

Research Method
The Web of Science databases published by Thomson Reuters were used for the study. The study was limited to the used of Social Science Citation Index (SSCI) and Arts and Humanities Citation Index (A&HCI). The data for the study were obtained by conducting an advanced search using Web of Science Category, WC=Information Science and Library Science, with articles as the document types, and all languages in terms of selection of language of publication of the journal articles. The period of the study was limited to 2010-2016 to depict current state of research globally and in Africa in particular. The data obtained were analyzed globally and by African countries based on the objectives of the study.

Results and Discussion
The results of the study are presented and discussed in this section based on the objectives of the study.

Ranking of LIS research and publications in Africa in global perspective
The results of the study showing ranking of LIS research and publications in Africa in global perspective are presented in table 1. The results revealed that Africa is not represented among the top 10 countries globally in LIS research and publication in all bibliometric indicators that are used in this study (publication counts and citation...
The results of the study showed that South Africa and Nigeria are the foremost countries in LIS research in term of publication counts. The findings of the study are consistent with previous studies that affirmed South Africa and Nigeria as the leading countries in LIS research in Africa (Aharony, 2011; Onyancha, 2007).

### Table 1: Top 20 globally ranked countries in LIS publications

<table>
<thead>
<tr>
<th>S N</th>
<th>Countries</th>
<th>Total Articles</th>
<th>Total Citations</th>
<th>Citations per Article</th>
<th>H-Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>USA</td>
<td>9,782</td>
<td>68,176</td>
<td>6.97</td>
<td>73</td>
</tr>
<tr>
<td>2</td>
<td>China</td>
<td>2,025</td>
<td>12,664</td>
<td>6.25</td>
<td>38</td>
</tr>
<tr>
<td>3</td>
<td>England</td>
<td>1,969</td>
<td>13,471</td>
<td>6.84</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>Spain</td>
<td>1,848</td>
<td>8,171</td>
<td>4.42</td>
<td>33</td>
</tr>
<tr>
<td>5</td>
<td>Canada</td>
<td>1,421</td>
<td>9,547</td>
<td>6.72</td>
<td>38</td>
</tr>
<tr>
<td>6</td>
<td>Australia</td>
<td>1,246</td>
<td>6,942</td>
<td>5.57</td>
<td>31</td>
</tr>
<tr>
<td>7</td>
<td>Germany</td>
<td>1,000</td>
<td>6,631</td>
<td>6.63</td>
<td>53</td>
</tr>
<tr>
<td>8</td>
<td>Taiwan</td>
<td>869</td>
<td>5,702</td>
<td>6.53</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>South Korea</td>
<td>866</td>
<td>5,455</td>
<td>6.30</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>Netherlands</td>
<td>832</td>
<td>8,721</td>
<td>10.48</td>
<td>39</td>
</tr>
<tr>
<td>11</td>
<td>Brazil</td>
<td>740</td>
<td>1,077</td>
<td>1.46</td>
<td>13</td>
</tr>
<tr>
<td>12</td>
<td>Italy</td>
<td>524</td>
<td>3,108</td>
<td>5.93</td>
<td>23</td>
</tr>
<tr>
<td>13</td>
<td>France</td>
<td>508</td>
<td>2,432</td>
<td>4.79</td>
<td>21</td>
</tr>
<tr>
<td>14</td>
<td>Belgium</td>
<td>410</td>
<td>2,848</td>
<td>6.95</td>
<td>24</td>
</tr>
<tr>
<td>15</td>
<td>Sweden</td>
<td>395</td>
<td>2,386</td>
<td>6.04</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>India</td>
<td>386</td>
<td>1,385</td>
<td>3.59</td>
<td>15</td>
</tr>
<tr>
<td>17</td>
<td>Singapore</td>
<td>381</td>
<td>2,570</td>
<td>6.75</td>
<td>24</td>
</tr>
<tr>
<td>18</td>
<td>Finland</td>
<td>374</td>
<td>2,730</td>
<td>7.30</td>
<td>24</td>
</tr>
<tr>
<td>19</td>
<td>South Africa</td>
<td>364</td>
<td>1,006</td>
<td>2.76</td>
<td>15</td>
</tr>
<tr>
<td>20</td>
<td>Denmark</td>
<td>295</td>
<td>2,041</td>
<td>6.92</td>
<td>22</td>
</tr>
</tbody>
</table>

### Ranking of LIS research and publication in Africa by country

The results of the study showing the rankings of LIS research by publication counts in Africa are shown in table 2. The results in table 2 revealed that South Africa and Nigeria are the foremost countries in LIS research in term of publication counts. The findings of the study are consistent with previous studies that affirmed South Africa and Nigeria as the leading countries in LIS research in Africa (Aharony, 2011; Onyancha, 2007).

### Table 2: Ranking of African countries by publication outputs

<table>
<thead>
<tr>
<th>S N</th>
<th>Country</th>
<th>Total Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>South Africa</td>
<td>364</td>
</tr>
<tr>
<td>2</td>
<td>Nigeria</td>
<td>165</td>
</tr>
<tr>
<td>3</td>
<td>Ghana</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>Kenya</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>Egypt</td>
<td>26</td>
</tr>
<tr>
<td>6</td>
<td>Tanzania</td>
<td>24</td>
</tr>
<tr>
<td>7</td>
<td>Botswana</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>Uganda</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>Tunisia</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>Malawi</td>
<td>11</td>
</tr>
<tr>
<td>11</td>
<td>Benin</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>Morocco</td>
<td>10</td>
</tr>
<tr>
<td>13</td>
<td>Zimbabwe</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>Algeria</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>Ethiopia</td>
<td>8</td>
</tr>
<tr>
<td>16</td>
<td>Senegal</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td>Namibia</td>
<td>6</td>
</tr>
<tr>
<td>18</td>
<td>Zambia</td>
<td>4</td>
</tr>
<tr>
<td>19</td>
<td>Mozambique</td>
<td>3</td>
</tr>
</tbody>
</table>

### Ranking of African countries by citation analysis

In table 3, African countries are ranked by citation analysis and the results revealed that South Africa is significantly leading other countries in the number of total citations (1006), followed by Nigeria (229), Kenya (106), Ghana (90) and Uganda (90). South Africa (15) is again leading in h-index, followed by Nigeria (7), while Tunisia (7.58) is the most ranked African country in terms of citations per article and is followed by Ethiopia (7.00).
Table 3: Ranking of African countries in LIS research by citation analyses

<table>
<thead>
<tr>
<th>SN</th>
<th>Countries</th>
<th>Total Citations</th>
<th>Citations per article</th>
<th>H-Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>South Africa</td>
<td>1,006</td>
<td>2.76</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>Nigeria</td>
<td>229</td>
<td>1.39</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>Ghana</td>
<td>90</td>
<td>2.37</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>Kenya</td>
<td>106</td>
<td>3.53</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>Egypt</td>
<td>89</td>
<td>3.42</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>Botswana</td>
<td>74</td>
<td>3.08</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>Tanzania</td>
<td>65</td>
<td>2.71</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>Uganda</td>
<td>90</td>
<td>4.29</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>Tunisia</td>
<td>91</td>
<td>7.58</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>Malawi</td>
<td>62</td>
<td>5.64</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>Benin</td>
<td>19</td>
<td>1.73</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>Morocco</td>
<td>32</td>
<td>3.20</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>Algeria</td>
<td>4</td>
<td>0.44</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>Zimbabwe</td>
<td>7</td>
<td>0.78</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>Ethiopia</td>
<td>56</td>
<td>7.00</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>Senegal</td>
<td>14</td>
<td>2.33</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>Namibia</td>
<td>6</td>
<td>1.00</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>Zambia</td>
<td>12</td>
<td>3.00</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>Mozambique</td>
<td>2</td>
<td>0.67</td>
<td>1</td>
</tr>
</tbody>
</table>

Trends in LIS research and publications in Africa, 2010-2016

The growth, progress and development of different fields of knowledge are usually determined by annual publication outputs in a given field (discipline) over a period of time. It was therefore pertinent to examine the trends in LIS research and publication in the study. The results of the study are presented in figure 1. The results showed an increasing trend in general or overall annual publication outputs in LIS research in Africa; but analysis by individual country using the three most ranked countries (South Africa, Nigeria and Ghana), showed declining and fluctuating trends. In South Africa, there was an increased in annual publication output until a decline was set in 2015-2016. The declining and fluctuating trend in annual publication outputs was apparent in Nigeria. Similar declining and fluctuating trend was also observed in annual publication output in Ghana. In the case of Nigeria, a recent study by Ani, Ngulube and Onyancha (2017) affirmed fluctuating and unpredictable trend in the annual publication output in LIS research in Nigeria.

References

Aharony, N. 2011. Library and Information Science Research Areas: A Content Analysis of Articles from the Top 10 Journals 2007-


A Rainbow at the Skyline  
after the Storm of Indicators for Ranking Scientists

Georgios Stoupas¹, Antonis Sidiropoulos², Antonia Gogoglou³, Dimitrios Katsaros⁴, Yannis Manolopoulos⁵

¹ grgstoupas@gmail.com, ² asidirop@it.teithe.gr  
Alexander Technological Educational Institute of Thessaloniki, Greece

³ {agogoglou, manolopo}@csd.auth.gr  
Aristotle University of Thessaloniki, Greece

⁴ dkatsr@inf.uth.gr  
University of Thessaly, Volos, Greece

Introduction

A plethora of bibliometric indices have been proposed to quantify scientific output. To deal with this storm of valuable indicators, the need arises for a classification scheme of scientists according to multiple evaluation metrics. In this work, we expand upon the concept of the skyline operator (Sidiropoulos, Gogoglou, Katsaros, & Manolopoulos, 2016), and introduce a new indicator, namely the Rainbow Ranking. For this study, we collected data from Microsoft Academic Search (MAS), and extracted full citation data starting from year 1950 up to 2015 for computer science scientists.

Rainbow-Ranking (RR-index)

The Rainbow-Ranking applies the skyline operator iteratively until all scientists are classified into a skyline level. Figure 1 shows a graphical representation of the skyline levels with two dimensions: citations per publication and the h-index. Every point in Figure 1 corresponds to a scientist. Each line connecting the points corresponds to a different skyline level. The x-axis represents ranking positions of each scientist according to their h-index, whereas on the y-axis the respective ranking positions according to citations per publication. Since this iterative procedure results into a plot with grouped curves as shown, we have named in the Rainbow Ranking.

For the rest of our experiments we select as dimensions of the skyline operator the h-index (Hirsch, 2005), the Perfectionism Index (Sidiropoulos, Katsaros, & Manolopoulos, 2015), and the A-index (Jin, Liang, Rousseau, & Egghe, 2007). Given a set of scientists A=X₁, the first call of skyline produces the first skyline level. We denote this first set of scientists as set S₁. Next, we compute set X₂=X₁-S₁, which contains the scientists that were not classified in the first skyline set S₁. For the set X₂ the skyline operator is applied once more and the result is the second skyline level (S₂). The process continues until all the scientists of the dataset are assigned a value.

To summarize the ranking levels into a single number metric, given a set of scientists A and a set of dimensions dims, we define the RR-index of a scientist a based on dims as follows:

\[
RR(dims) = \frac{\left| \text{levels}(a, dims) \right| + \left| \text{ties}(a, dims) \right|}{2|A|}
\]

|A| is the total number of scientists, |levels(a,dims)| is the number of scientists ranked at higher skyline levels than scientist a based on dimensions dims. Level 1 is considered higher than level 2. |ties(a,dims)| is the number of scientists who are ranked at the same level with scientist a, excluding scientist a. Thus, the following holds for the RR-index: 0 < RR(dims) ≤ 100.

Results

The following table illustrates the RR-based ranking. The first skyline level is occupied by scientists who can be grouped into two groups; one group (in italics) are those who have worked in core computer science (e.g., networking, compilers, databases), and the second group are those who have worked in computational biology. In the second skyline level, we mainly encounter core computer scientists (in italics), but also a political scientist and economist (Simon Herbert); the others are computational biologists.
### Table 1. First two skyline levels' members

<table>
<thead>
<tr>
<th>Name</th>
<th>C</th>
<th>P</th>
<th>C/P</th>
<th>h-index</th>
<th>A-index</th>
<th>PI</th>
<th>RR</th>
<th>Skyline level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shenker Scott</td>
<td>38557</td>
<td>473</td>
<td>81.52</td>
<td>90</td>
<td>361.02</td>
<td>12187</td>
<td>99.99</td>
<td>1</td>
</tr>
<tr>
<td>Foster Ian</td>
<td>39052</td>
<td>730</td>
<td>53.50</td>
<td>87</td>
<td>365.48</td>
<td>-9320</td>
<td>99.99</td>
<td>1</td>
</tr>
<tr>
<td>Ullman Jeffrey</td>
<td>38019</td>
<td>445</td>
<td>85.44</td>
<td>82</td>
<td>394.98</td>
<td>14977</td>
<td>99.99</td>
<td>1</td>
</tr>
<tr>
<td>Haussler David</td>
<td>27799</td>
<td>320</td>
<td>86.87</td>
<td>78</td>
<td>305.29</td>
<td>15007</td>
<td>99.99</td>
<td>1</td>
</tr>
<tr>
<td>Tibshirani Robert</td>
<td>47661</td>
<td>344</td>
<td>138.55</td>
<td>69</td>
<td>636.06</td>
<td>33447</td>
<td>99.99</td>
<td>1</td>
</tr>
<tr>
<td>Miller Webb</td>
<td>54262</td>
<td>532</td>
<td>102.00</td>
<td>42</td>
<td>1272.76</td>
<td>35446</td>
<td>99.99</td>
<td>1</td>
</tr>
<tr>
<td>Higgins Desmond</td>
<td>41527</td>
<td>190</td>
<td>218.56</td>
<td>21</td>
<td>1974.43</td>
<td>38419</td>
<td>99.99</td>
<td>1</td>
</tr>
<tr>
<td>Lipman David</td>
<td>48638</td>
<td>97</td>
<td>501.42</td>
<td>19</td>
<td>2453.42</td>
<td>45970</td>
<td>99.99</td>
<td>1</td>
</tr>
<tr>
<td>Altschul Stephen</td>
<td>46730</td>
<td>78</td>
<td>599.10</td>
<td>85</td>
<td>365.48</td>
<td>12187</td>
<td>99.99</td>
<td>1</td>
</tr>
<tr>
<td>Gish Warren</td>
<td>26065</td>
<td>33</td>
<td>789.85</td>
<td>2894.11</td>
<td>25930</td>
<td>99.99</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Thompson Julie</td>
<td>36441</td>
<td>450</td>
<td>80.98</td>
<td>4552.50</td>
<td>32969</td>
<td>99.99</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Gibson Toby</td>
<td>36329</td>
<td>427</td>
<td>85.08</td>
<td>4538.63</td>
<td>33041</td>
<td>99.99</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Zhang Jinghui</td>
<td>28638</td>
<td>94</td>
<td>304.66</td>
<td>5727.20</td>
<td>28218</td>
<td>99.99</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Garcia-Molina Hector</td>
<td>25743</td>
<td>578</td>
<td>44.54</td>
<td>205.88</td>
<td>-9173</td>
<td>99.98</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Estrin Deborah</td>
<td>34706</td>
<td>446</td>
<td>77.82</td>
<td>344.86</td>
<td>11246</td>
<td>99.98</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Culler David</td>
<td>27360</td>
<td>363</td>
<td>75.37</td>
<td>296.17</td>
<td>11267</td>
<td>99.98</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Simon Herbert</td>
<td>31620</td>
<td>1194</td>
<td>26.48</td>
<td>389.40</td>
<td>-46680</td>
<td>99.98</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Lander Eric</td>
<td>42201</td>
<td>430</td>
<td>98.14</td>
<td>612.10</td>
<td>22369</td>
<td>99.98</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Rivest Ronald</td>
<td>38336</td>
<td>294</td>
<td>130.39</td>
<td>618.14</td>
<td>28012</td>
<td>99.98</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Vapnik Vladimir</td>
<td>31324</td>
<td>123</td>
<td>254.67</td>
<td>618.14</td>
<td>30099</td>
<td>99.98</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Leiserson Charles</td>
<td>23147</td>
<td>155</td>
<td>149.34</td>
<td>627.36</td>
<td>20159</td>
<td>99.98</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Myers Eugene</td>
<td>32210</td>
<td>286</td>
<td>112.62</td>
<td>954.42</td>
<td>24950</td>
<td>99.98</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Cormen Thomas</td>
<td>16707</td>
<td>50</td>
<td>334.14</td>
<td>1189.57</td>
<td>16399</td>
<td>99.98</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Shannon Claude</td>
<td>13554</td>
<td>32</td>
<td>423.56</td>
<td>1935.57</td>
<td>13428</td>
<td>99.98</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Woods Richard</td>
<td>11642</td>
<td>41</td>
<td>283.95</td>
<td>1940.33</td>
<td>11468</td>
<td>99.98</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Schaffer Alejandro</td>
<td>24096</td>
<td>42</td>
<td>573.71</td>
<td>4818.20</td>
<td>23936</td>
<td>99.98</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

### Summary

This article addresses the following problem: “Given a set of bibliometric indicators, selected in any algorithmic way, can we successively rank scientists into layers based on these indicators, such that the scientists in each layer outperform those of the lower layers according to (at least one) indicator?” We employed the skyline and iteratively applied it to scientists, thus designing the **Rainbow Ranking** indicator. We evaluated it against computer scientists and showed intuitive results.

### References


An Evolutionary Analysis of Research Integrity Based on Multidimensional Informetrics

Chen Yu1  Li Chenying 2  Zhao Yong3

1 chenyu2016@cau.edu.cn
2 licy@cau.edu.cn
3 zhaoyong@cau.edu.cn

Information Research Center of Library, China Agricultural University, No.2, Yuanmingyuan West Road, Haidian District, Beijing (China)

Introduction

The rapid development of information technology has driven the media dissemination of information on the speed and scope, so the frequent exposure of the academic misconduct caused great concern for research integrity in today's society and seriously damaged the public trust in science. Since 1952, The Integrity of Science published in Nature anonymously, The research integrity has received widespread concern around the world. At least four thousand related papers from China have been published in academic journals. The number of papers in recent years is growing fast and the problem of research integrity has become a focus of academic circles.

From the perspective of the development of science and technology, and on the basis of the development and evolution of the concept of research integrity, the paper applies bibliometric, policy text analysis and content analysis methods, aiming at providing a reference for understanding the evolution of research research in China.

Data and Methods

Data collection

WOS was used as the foreign source of data analysis. With "TI=("research integrity") or ("scien* integrity") or ("academic* integrity") or ("research misconduct") or ("scien* misconduct") or ("academic* dishonest") or ("academic* anomie") or ("research morality") "for the search strategies. processed a total of 1019 related documents. This part of the documents are used as a sample of research direction content analysis.

CNKI was used as the domestic source of the data analysis. With "TI=科研诚信 or 学术诚信 or 科研不端 or 学术不端 or 科研道德 or 学术道德 or 科研失范 or 学术失范 or 科研腐败 or 学术腐败 or 科研造假 or 学术造假 or 学术抄袭 or 学术剽窃" for the search strategies. finally received 1070 related literature.

The study collected 37 relevant policy texts published on 19 government or research agency websites in 12 countries This part of the data were used as a sample of the policy text analysis.

Data analysis

In the study, the paper mainly adopted the standardized processing technology of the literature title data, standardly processed the proposed keyword field, economically analyzed these fields to obtain frequency or co-occurrence of the frequency of statistical data, then applied Ucinet social network analysis tools for keyword co-occurrence analysis and keyword center degree analysis, and finally used Netdraw for visualization. The specific research hotspot analysis process shown in Figure 1.

Figure 1. The Division process of Research integrity's hotspots

Results and discussion

An overview

The distribution of the literature can reflect the level of academic research in a certain extent (Sun Yusheng et al. 2013), The speed of development and the degree of attention, revealing the longitudinal law of the study (Liu Guifeng, 2011). Prior to 2016, a total of 2089 articles were published in the core journals of WOS and CNKI. Figure 1 shows the number of articles in WOS and CNKI for each period. As shown in Figure 1, the X-axis represents the age distribution, and the Y-axis represents the number of documents. The two columns in each period show the number of WOS and CNKI articles.

Figure 2. Distribution of the number of articles in domestic and foreign

From Figure 1 we can found that the scientific integrity at home and abroad is in a upward trend on the whole. As for the research of scientific integrity, foreign research started earlier, and the number of research results were more than that of...
China. However, after 2007, the research results in China started to overnumber those in foreign countries.

**The evolution of policy and conceptual terms**

In 1942 the American sociologist RK Merton first proposed the concept and basic scientific norms of modern science in On Science and Democracy, and in the 1980s it began to evolve from academic issues to social and government public policies. By the end of the 1980s, the "Schoen case" of USA, "Huang Yuxi paper fraud" of South Korea and other international scientific misconduct occurred frequently. The scientific integrity issue has upgraded to the focus of the international community. This reflects the evolutionary path of scientific integrity from the bud, start to the development and internationalization.

In 1990s, China really began to pay attention to scientific integrity. Over the past three years, about 100 theoretical research literature can be retrieved each year through the application of "scientific integrity" as the keyword to search article from CNKI. In particular, in 2011, the China Association for Science and Technology, the Ministry of Education jointly launched the postgraduate scientific morality and style of study to promote educational activities, academic ethics and scientific integrity education work to a new height.

![Figure 3. development memorabilia of Research Integrity](image)

**Figure 3. development memorabilia of Research Integrity**

![Figure 4. The first occurrence of the concept of research integrity in academic papers](image)

**Figure 4. The first occurrence of the concept of research integrity in academic papers**

**Combing the evolution of research focus**

Keywords can summarize the core content of the paper. In order to explore the research content of scientific integrity, this paper extracted all the keywords of the research literature, pre-processed with Bibstats, and sorted the word frequency of the keywords. Research integrity's hotspots development evolution is divided into the following three stages:
- Research on the Construction of Academic Moral Spirit;
- Research on Academic evaluation and academic management;
- Research on research integrity education in colleges and universities

**Conclusion**

Through the above analysis, it is found that the progress of scientific integrity research in China has lagged far behind the developed countries. The initial research of integrity research mainly focused on the explanatory study of the academic moral. The number of research papers is small and the content is limited to criticizing academic moral misconduct or cultivating the importance of academic ethics and other surface level. After the 2000, research in the mid-term research inclined to descriptive research. It began to explore based on the case and gradually conducted in-depth, systematic research on the theme of integrity research, meanwhile discussing the problems and governance methods as well as researching the governance of all aspects of the research misconduct. The paper from 2008 to now is more inclined to exploratory research, and the direction of research is more in-depth, concerned with the guidance research of the policy establishment, and more focused on academic management research. Management of research integrity transforms from "the problem re-governance" into predictive "take precautions". In the past two years, the research on the field of scientific integrity has been carried out from the perspective of reality and theory.

**Acknowledgments**

We acknowledge support from the Institute of Scientific and Technical Information of China (2016JP022 — Related research of scientific integrity and scientific ethics) and (2015JP006 — Research on the scheme and path of establishing scientific credit management system in the management system of technology plan).

**References**


Do under-cited influential sleeping beauties exist?

Xiaojun Hu\textsuperscript{1} Ronald Rousseau\textsuperscript{2}

\textsuperscript{1}xjhu@zju.edu.cn
Medical Information Center, Zhejiang University School of Medicine, Hangzhou 310058 (China)

\textsuperscript{2}ronald.rousseau@kuleuven.be
KU Leuven, B-3000 Leuven (Belgium) \& University of Antwerp (UA), Faculty of Social Sciences, B-2000 Antwerp, (Belgium)

Abstract
In this study we investigate if sleeping beauties, also known as articles suffering delayed recognition, can at the same time be under-cited influential articles. Theoretically these two types of articles are independent, in the sense that being a sleeping beauty depends on the number and time distribution of received citations, while being an under-cited influential article depends only partially on the number of received (first generation) citations, and much more on second and third citation generations. Among 49 sleeping beauties we found 13 that are also under-cited influential.

Conference Topic
Citation Analysis

Introduction
In diachronous studies and depending on their citation curves or ageing profiles articles can be categorized in different ways. In (Smith Aversa, 1985) the author makes, for highly cited papers, a distinction between delayed rise in received citations combined with a slow decline, on the one hand, and early rise combined with a rapid decline, on the other. More recently, Costas et al. (2010) classify documents into three general types: delayed documents, which receive the main part of their citations later than normal documents; flashes-in-the-pan (van Dalen \& Henkens, 2005), which receive citations immediately after their publication but are not cited or influential in the long term; and normal documents, documents with a typical distribution of citations over time. The delayed documents include articles known as sleeping beauties (van Raan, 2004), articles suffering from delayed or late recognition, being premature discoveries, being ahead of one's time, suffering from Mendel's syndrome or being late bloomers (all synonyms or near-synonyms, describing the same phenomenon).

Recently, and especially in China, sleeping beauties attracted a lot of attention (Ke et al., 2015; Li, 2014; Li \& Shi, 2016; Li \& Ye, 2016; Du \& Wu, 2016), leading among others to “all-elements-sleeping-beauties” (Li \& Ye, 2012), i.e. articles that at first receive some attention, then go through a period in which they are hardly noticed, followed by a period of being highly cited, awakened by a prince, so to speak. Figure 1 shows the citation curve of (Romans, 1986), a typical sleeping beauty studied by van Raan (2004). It received no citations for nine years and then received more than 100 citations over the next six years. After a peak (it received 32 citations in the year 1999) its citation curve shows a gradual and irregular decline similar to that of ‘normally cited’ articles.
A different type of articles: under-cited influential articles

All the types of articles mentioned above were based on the – direct – citation distribution. Yet, recently Hu and Rousseau (2016, 2017) introduced another type of articles in which first, second and third generation citations play a role, leading to the sparking index (Hu & Rousseau, 2017). These articles are referred to as under-cited influential articles. Such articles are characterized by three properties: (1) they are reasonably well-cited (a basic requirement to be influential); (2) citations of citations (second generation citations) are rather high, so that the original one is influential in an indirect way (a more refined token of influence); (3) given condition two, these articles received fewer citations than expected (being under-cited). Each of these three requirements is operationalized leading to a two formulae referred to as sparking indices. There are two types of sparking indices depending on the number of received citations: a sparking index on the 1% level, denoted as S1 (for articles that are cited at least 201 times) and a sparking index on the 10% level, denoted as S10 (for articles that are cited at least 21 times but less than 201). Details are provided in (Hu & Rousseau, 2016, 2017). These two articles also show that such under-cited influential articles are neither common, nor extremely rare. Many articles written by Nobel Prize winners, but also other, including recent articles, fall into the group of under-cited influential articles.

Before continuing our investigation we want to clarify the terminology of under-cited influential articles. It is clear that these articles are influential: they are reasonably well cited and, through citing articles, their influence on the development of the ideas that started with them is even higher. Many of them, like the Nobel prize winning articles studied in (Hu & Rousseau, 2017), can even be described as foundational articles. These articles are also referred to as under-cited. By this we do not mean that they are treated unfairly. We just want to stress that many articles that built on them received more citations. One reason could be that these articles added new results that incorporated the results of the ‘under-cited’ one, so that later investigators do not feel the need to refer to the original.

Clearly sleeping beauties and under-cited influential articles differ in many aspects except that in short-term evaluations based on bibliometric indicators, they both are under-valued. Sleeping beauties usually begin with a long period of few citations and then direct citations (first generation) go up. Under-cited influential articles usually are rather well-cited but
thenumber of received citations does not fully reflect their influence on their field. Probably only expert peer reviewers can see their importance. These considerations led us to the following research question.

**Are there sleeping beauties which are also under-cited influential articles?**

We looked through a set of articles dealing with delayed recognition or sleeping beauties and determined 49 articles considered to be sleeping beauties, restricting then to 1950 or later and included in the SCI-E. As a universally accepted, precise definition of a sleeping beauty does not exist, we do not discuss if these articles are really sleeping beauties. We just accept the opinion of the authors who studied these articles. For all these sleeping beauties we checked if they are also under-cited influential articles. Citations were collected and calculations made in February 2017. Bibliographic information on these sleeping beauties is given in Table 2 (Appendix), while the articles discussing these sleeping beauties can be found in Table 3 (Appendix). Results of our investigations are shown in Table 1. We found 13 sleeping beauties of the 49 checked that are also under-cited influential articles. This is 27%. Of course numbers are too small to accord any meaning to this percentage.

**Table 1. Articles that are sleeping beauties and moreover also under-cited influential**

<table>
<thead>
<tr>
<th>First author [number in Table 2]</th>
<th>PY</th>
<th>Discussed as sleeping beauty by the following author(s); see Table 3</th>
<th>Received citations</th>
<th>Sparking index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leakey [2]</td>
<td>1964</td>
<td>Tobias</td>
<td>310</td>
<td>$S_1 = 358$</td>
</tr>
<tr>
<td>Scott [3]</td>
<td>1973</td>
<td>Ohba &amp; Nakao</td>
<td>320</td>
<td>$S_1 = 343$</td>
</tr>
<tr>
<td>De Rujula [4]</td>
<td>1977</td>
<td>Li</td>
<td>191</td>
<td>$S_{10} = 210$</td>
</tr>
<tr>
<td>Gitter [8]</td>
<td>1980</td>
<td>Glänzel et al.</td>
<td>88</td>
<td>$S_{10} = 176$</td>
</tr>
<tr>
<td>Lis [9]</td>
<td>1980</td>
<td>Glänzel et al.</td>
<td>763</td>
<td>$S_1 = 966$</td>
</tr>
<tr>
<td>Ogino [10]</td>
<td>1980</td>
<td>Glänzel et al./Braun et al.</td>
<td>483</td>
<td>$S_1 = 1078$</td>
</tr>
<tr>
<td>Waite [12]</td>
<td>1980</td>
<td>Braun et al.</td>
<td>164</td>
<td>$S_{10} = 176$</td>
</tr>
<tr>
<td>Romans [13]</td>
<td>1986</td>
<td>Van Raan</td>
<td>347</td>
<td>$S_1 = 723$</td>
</tr>
</tbody>
</table>

**Characteristics of articles which are sleeping beauties as well as under-cited influential**

The article by Romans (1986) turned out to be also under-cited influential according to our definition. This is illustrated in Figures 2-3 where, besides the original citation curve we also show the curve of the citing curves with the second most citations (in yearly distribution and in cumulative number). As the top 1% (rounded up) citing articles consists of four articles we show the second one among these (referred to as the median citing article). This median citing article, published in 1999, received already in the year 2000 more citations that Romans’ (Figure 2) and the next year its cumulative total citations was higher than Romans’ (Figure 3). This is just an illustration as such an intersection is neither necessary nor sufficient to be an under-cited influential article.
Discussions and conclusions

We note the following caveat: the definitions of sleeping beauties and under-cited influential articles depend on some ad hoc rules. This leads to the question: do these 13 articles, or most of these, really belong to the two types mentioned, or would a slight change in definition to one of the two types make a consequential difference?

Assume that we have a strict definition of a sleeping beauty, then we wonder after how many years is it clear that an article is a sleeping beauty? Is it, at that moment, also possible to say that this sleeping beauty is moreover under-cited influential (for which a strict definition exists), or is it then usually too early, as one needs three generations of citing articles and it took quite some time before the first generation was ‘large enough’ and hence does not easily leads to many second and third generation citations. Assuming that a ‘recent’ sleeping beauty is not under-cited influential is it then possibly to predict (i.e., determine a probability) if it will ever be?
We conclude that we have shown that there do exist articles that are sleeping beauties and are at the same time under-cited influential, Romans’ (1986) being a case in point.

Acknowledgments
This work was supported by the National Natural Science Foundation of China Grant 71573225. The authors would like to thank Zhang Yuning and Hu Xiaoyue for their help in data collection.

References
Li, J. (2014). Citation curves of "all-elements-sleeping-beauties": "flash in the pan" first and then "delayed recognition". Scientometrics, 100(2), 595-601.

Appendix
Table 2. List of sleeping beauties studied in this contribution. All articles are published in 1950 or later and are included in the SCI-E. Articles that discuss these sleeping beauties are listed in Table 3.


**Table 3. Articles discussing sleeping beauties studied in our investigation.**


Li, J. (2014). Citation curves of "all-elements-sleeping-beauties": "flash in the pan" first and then "delayed recognition". *Scientometrics*, 100(2), 595-601.


Which Drives Which? The Causal Relationship between Number of Editorial Board Members and Scientific Output of Universities in the Chemistry Field: a Granger Causality Test

Xing Wang\textsuperscript{1,2}

\textsuperscript{1} wangxing@sjtu.edu.cn
Shanxi University of Finance & Economics, Information Management School, 696 Wucheng Road, 030006 Taiyuan (China)

\textsuperscript{2} wangxing830914@gmail.com
Shanghai Jiao Tong University, Graduate School of Education, 800 Dongchuan Road, 200240 Shanghai (China)

Introduction
Editorial boards of scholarly journals are important to the entire academic world, and there may be some relationship between the number of editorial board members and the scientific output of universities. Several studies have examined the correlation between the number of editorial board members and scientific output of universities in some subjects (Braun, Dióspatonyi, Zádor, & Zsindely, 2007; Burgess & Shaw, 2010; Frey & Rost, 2010). However, there has been a lack of study about the causality in this relationship. It may be more interesting to determine whether such a causal relationship exists between the number of editorial board members and the scientific output of universities and, if so, which drives which.

In this study, we used time-series data and the Granger causality test to explore the causal relationship between the number of editorial board members and the number of articles of top 14 universities in the chemistry field.

Data and Method
We collected time-series data for the Granger causality test using two variables: the number of editorial board members and the number of articles published per university. We selected the following nine top journals as sample journals for this analysis: Journal of the American Chemical Society, Angewandte Chemie International Edition, Chemical Reviews, Accounts of Chemical Research, Analytical Chemistry, Biochemistry, Chemistry of Materials, Inorganic Chemistry, and Journal of Organic Chemistry. We chose the period from 1998 to 2014.

We used the Shanghai ranking’s top 20 universities in chemistry as our sample universities for the Granger causality test. First, we recorded and calculated the number of editorial board members over the years of 20 universities from 1998 to 2014 at the nine journals. Next, we artificially set a threshold: only universities with no less than five editorial board members for at least one year were considered qualified for our Granger causality test. Fourteen universities constituted the final sample for this analysis.

Using the advanced search function of Web of Science, we obtained the number of articles published in the nine journals each year from 1998 to 2014 at 14 universities. The main statistical technique used in this study was Granger causality test model which can be written as follows:

\[
Y_t = \alpha_0 + \sum_{i=1}^{m} \alpha_i Y_{t-i} + \mu_t \tag{1}
\]

\[
Y_t = \alpha_0 + \sum_{i=1}^{m} \alpha_i Y_{t-i} + \sum_{j=1}^{m} \beta_j X_{t-j} + \mu_t \tag{2}
\]

Results
The Granger cause test results showed that there was unidirectional causality for only three universities which are UCLA, ETH-Zurich, and University of Pennsylvania. However, there was no significant causal relationship in either direction for the other eleven universities (Table 1). The regression equations of UCLA, ETH-Zurich, and the University of Pennsylvania based on the Granger causality test model (2) are shown in Table 2.

Although the Granger causality test results suggested unidirectional causality in the above three universities, the established regression equations of these three universities based on the Granger causality test model were contradictory with respect to meaning. For example, the coefficient of $X_{t-1}$ was significantly negative ($p < 0.05$) in the equation of lag phase 1 of UCLA, and the coefficient of $X_{t-1}$ was also significantly negative ($p < 0.05$) in the equation of the lag phase 2 of UCLA. Similar issues occurred in the situation of ETH-Zurich and the University of Pennsylvania.
Based on the above results, the causal relationship between the number of editorial board members and the number of articles of top 14 universities in chemistry was not obvious overall.

Table 1. Granger cause test between the number of editorial board members and the number of articles of universities (1998-2014).

<table>
<thead>
<tr>
<th>R</th>
<th>University</th>
<th>EB→PUB</th>
<th>PUB→EB</th>
<th>Lag</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UC-Berkeley</td>
<td>1.669</td>
<td>1.975</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.316)</td>
<td>(0.263)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Harvard Univ</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>3</td>
<td>Stanford Univ</td>
<td>1.697</td>
<td>1.952</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.215)</td>
<td>(0.186)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Northwestern Univ</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>5</td>
<td>MIT</td>
<td>1.269</td>
<td>2.098</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.323)</td>
<td>(0.174)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Caltech</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>8</td>
<td>ETH-Zurich</td>
<td>0.704</td>
<td>5.162**</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.518)</td>
<td>(0.029)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.693</td>
<td>4.443**</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.585)</td>
<td>(0.048)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.507</td>
<td>5.283*</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.737)</td>
<td>(0.068)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Kyoto Univ</td>
<td>1.384</td>
<td>1.246</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.380)</td>
<td>(0.418)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>UCLA</td>
<td>5.644**</td>
<td>1.529</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.034)</td>
<td>(0.238)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.693**</td>
<td>2.211</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.037)</td>
<td>(0.160)</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Univ of Pennsylvania</td>
<td>0.812</td>
<td>49.647***</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.578)</td>
<td>(0.001)</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Yale Univ</td>
<td>0.001</td>
<td>0.621</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.971)</td>
<td>(0.445)</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>UC-Santa Barbara</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>14</td>
<td>UC-San Diego</td>
<td>2.017</td>
<td>0.282</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.257)</td>
<td>(0.876)</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Univ Tokyo</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
</tbody>
</table>

Note. EB, PUB represents respectively the number of editorial board members and the number of articles. Column 1 is the 2014 Shanghai ranking of the selected universities in chemistry. Column 5 is the lag phase. For the three universities that had a significant causal relationship between the two variables, the lag phases having a significant causal relationship were all provided. For universities with no detected causal relations during lag phase 1 to 4, we only present an optimal lag phase based on the Akaike Information Criterion. "**"] indicate that the EB and Pub series are not stationary or not co-integrated.

Table 2. Regression equation based on Granger causality test model of the three universities where Causal relationships were detected.

<table>
<thead>
<tr>
<th>Lag</th>
<th>Regression equation based on Granger causality test model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( Y_t = 91.979 + 0.379 Y_{t-1} - 4.508 X_{t-1} )</td>
</tr>
<tr>
<td>2</td>
<td>( Y_t = 152.597 + 0.150 Y_{t-1} - 0.078 Y_{t-2} )</td>
</tr>
<tr>
<td></td>
<td>(-4.405 X_{t-1} - 4.053 X_{t-2} )</td>
</tr>
<tr>
<td></td>
<td>Part B: ETH-Zurich</td>
</tr>
<tr>
<td></td>
<td>( X_t = 4.700 + 0.622 X_{t-1} - 0.374 X_{t-2} )</td>
</tr>
<tr>
<td></td>
<td>(+0.054 Y_{t-1} - 0.062 Y_{t-2} )</td>
</tr>
<tr>
<td>3</td>
<td>( X_t = 4.583 + 0.902 X_{t-1} - 0.529 X_{t-2} )</td>
</tr>
<tr>
<td></td>
<td>(-0.168 X_{t-3} + 0.050 Y_{t-1} - 0.085 Y_{t-2} )</td>
</tr>
<tr>
<td></td>
<td>(+0.033 Y_{t-1} )</td>
</tr>
<tr>
<td></td>
<td>Part C: University of Pennsylvania</td>
</tr>
<tr>
<td></td>
<td>( X_t = 8.617 + 0.533 X_{t-1} - 0.185 X_{t-2} )</td>
</tr>
<tr>
<td></td>
<td>(-0.334 X_{t-3} + 0.501 X_{t-4} + 0.039 Y_{t-1} )</td>
</tr>
<tr>
<td></td>
<td>(-0.060 Y_{t-2} - 0.005 Y_{t-3} + 0.019 Y_{t-4} )</td>
</tr>
<tr>
<td>4</td>
<td>( X_t = 5.617 + 0.304 X_{t-1} + 0.625 X_{t-2} )</td>
</tr>
<tr>
<td></td>
<td>(-0.069 X_{t-3} + 0.777 X_{t-4} + 0.007 Y_{t-1} )</td>
</tr>
<tr>
<td></td>
<td>(-0.025 Y_{t-2} - 0.027 Y_{t-3} - 0.004 Y_{t-4} )</td>
</tr>
</tbody>
</table>

Note. X, Y represents respectively the number of editorial board members and the number of articles. Bold and italics denote the associated p-values lower than 5%.

Conclusion

Our Granger causality test results suggest that the causal relationship between the number of editorial board members and the number of articles of top 14 universities in chemistry is not obvious overall. Combining with the interviews of some editorial board members from the above nine journals, we consider that the causal relationship between the number of editorial board members and the number of articles of universities may not be ‘rigid’.

References


Abstract
The vector representation of documents is important for literature aggregation, clustering and classification. Co-occurrence latent semantic vector space model (CLSVSM) based on the binary co-occurrence information, exploited the latent semantic relations between terms and documents. Comparing with the vector space model (VSM), it can improve the accuracy of text clustering. To better utilize the latent semantic information through CLSVSM, this paper introduces the ternary co-occurrence information. Our goal in this survey is to offers a ternary co-occurrence layer matrix representation, to give a calculation method of ternary co-occurrence frequency and relative strength and to show the effect of the binary and ternary weighted CLSVSM. Finally, we use Chinese data sets and English data sets respectively to compare the two models. The results show that the new model is equal to binary CLSVSM in English data sets, but has an outstanding performance in Chinese data sets.

Key words
Co-occurrence analysis, Ternary co-occurrence, Co-word matrix, Weighted CLSVSM, Clustering

Conference Topic
Co-occurrence analysis

Introduction
In the era of big data, literature resources are very rich, but the resource overload tends to bring some questions, such as declining of the retrieval accuracy and increasing of the cost. Resource aggregation is an effective way to solve the question of too many resources. Literature aggregation can be realized by the clustering of the feature vector of literature; its result is related to whether the semantic information is fully extracted, whether the vector representation is reasonable, and whether the clustering method is effective, etc. It is very important to use the semantic information to express the vector fully. Literature representation is derived from the research of the text representation. The text representation methods include set theory, algebra, probability statistics, graph theory and so on. The most basic representation is the vector space model (VSM) (Salton, 1997), this model assumes independency between the vocabulary terms and ignores all the conceptual relations between terms that potentially exist; The Generalized vector space model (GVSM) (Wong, Ziarko&Wong, 1985) excavates the co-occurrence information of words, but it cannot extract the semantic information fully. The semantic vector space model [4] (SVSM) (Song, Liang&Soon, 2014), excessive relies on ontology or the language database. The co-occurrence latent semantic vector space model based on the binary information (CLSVSM) (Niu&Qiu, 2014) is used to proceed literature aggregation. The result shows that the model is better than the VSM. But we find that researching the binary information only is not comprehensive. In this paper, we introduce the ternary information.

In CLSVSM, the extraction of latent semantic information is based on co-occurrence analysis. Co-occurrence analysis is a method to quantify the co-occurrence information in various
information carriers. Its methods include co-cited analysis, bibliographic coupling analysis and co-word analysis (Pang, Fang & Yang, 2012). The research of this paper mainly starts from the co-word analysis, and digs the latent semantic information between the terms in the digital literature. The co-word analysis was proposed by Callon, Courtial and Turner et al (1983). It is a method of content analysis, and it is mainly used to analyze the phenomenon of professional terms that can reveal a subject or a research direction in a field occurring in one document. Take the frequency of a group of terms in one document as the foundation for hierarchical clustering and find the relationship between terms (Feng & Leng, 2006; Lu, Jing & Fan, 2016; Chen, Chen & Wu, 2016). Co-word analysis can be widely used in hot-spot analysis researching (Xie, Liang & Wang, 2005; Ma & Wang, 2007; Su, 2009; Li & Dong, 2010), discipline structure revealed (Li & Wang, 2011; Yang & Cui, 2011), information retrieval (Rokaya, Atlam & Fuketa, 2011) and other fields.

At present, there is less research of ternary relations. Dong, Liu and Jiang discussed the principle of the triad census method and applied it in the research on the structure of citation network (2010). Li proposed the concept of three-word sequential co-occurrence pair and the application of reflected terms and tight ring (2013). Hu and Chen proposed that the conceptual knowledge network based on keyword co-occurrence can be divided into different layers in terms of the k-core value of nodes, and introduced the theory of triadic closure as the basic unit of network structure analyses (2014). In the research of co-word, Leng, Wang and Li (2011) used the bit vector orthogonal method of DLG (direct large item-set generation) to give the co-word operation method, calculated the stability of ternary co-occurrence and its influence of the results, and gave the results comparison of binary and ternary co-word. Lastly, the author came to a conclusion: The binary co-word results reflected the relationship between research topics, which mainly including the keywords, and the result of the ternary co-word reflected the research area and questions of the phrase collocation. This study found the result of ternary word has some practical significance and it is very important to research the ternary information.

In this paper, the co-occurrence model is based on the co-word analysis, but it can apply for all co-occurrence analysis. This article is uniformly referred to as co-occurrence, for example, we present the co-occurrence of two (three) characteristic terms as binary (ternary) co-occurrence. The organization of the paper is as follows: In section 2, a brief review of the co-occurrence latent semantic vector space model based on binary information is given. The ternary co-occurrence layer matrix representation, ternary co-occurrence frequency, relative strength of ternary co-occurrence calculation and the weighted CLSVSM formation are presented in section 3. Further, experimental results and discussions to support the algorithm can be seen in section 4. Conclusions are derived in section 5.

**Co-occurrence latent semantic vector space model based on the binary information**

Let $A$ be a document - term matrix. Suppose our document collection contains $n$ documents and $m$ unique terms. The matrix $A$ will have $n$ rows (one row for each document) and $m$ columns (one column for each unique term in the vocabulary). Let $t_j$ be the $j$-th term in the vocabulary and let $d_i$ be the $i$-th document in the collection. The $i$-th row in $A$ is the row vector $a_i$, and the $j$-th column in is the column vector $a_j$. The row vector $a_i$ contains $m$ elements, one element for each term, and the column vector $a_j$ contains $n$ elements, one element for each document. Suppose $A$ is a simple matrix of frequencies. The element $a_{ij}$ in $A$ is the frequency of the $j$-th term in the $i$-th document. If the $j$-th term is
belong to the document $d_i$, then $a_{ij}$ is equal to 1, otherwise it is 0. The matrix is sparse and we define this matrix as Boolean document-term matrix.

Let $C$ be the co-occurrence matrix and let $B$ be the co-occurrence relative strength between terms. They are expressed as follow:

$$C = A^T \cdot A = \left(c_{ij}\right)_{m \times m}$$

$$B = \left(b_{ij}\right)_{m \times m} = \text{diag}\left(\frac{1}{\sqrt{c_{11}}}, \frac{1}{\sqrt{c_{22}}}, \ldots, \frac{1}{\sqrt{c_{mm}}}\right)A^T \cdot \text{diag}\left(\frac{1}{\sqrt{c_{11}}}, \frac{1}{\sqrt{c_{22}}}, \ldots, \frac{1}{\sqrt{c_{mm}}}\right)$$

$$b_{ij} = \frac{c_{ij}}{\sqrt{c_{ii}c_{jj}}}, i, j = 1, 2, \ldots, m$$

The CLSVSM is defined as follow:

$$\phi: d_i \rightarrow \phi(d_i) = (q_{i1}, q_{i2}, q_{i3}, \ldots, q_{im})^T \in R^D$$

$$q_{ij} = \begin{cases} 
1, & a_{ij} = 1 \\
\max_{l \in I_{l1}} \{b_{lj}\}, & a_{ij} = 0, \max_{l \in I_{l1}} \{b_{lj}\} \neq 0 \\
0, & \text{others.}
\end{cases}$$

Where $I_{l1} = \{j | a_{ij} = 1\}$ and $q_{ij} \in [0,1]$

The new document-term matrix is defined as:

$$Q = \begin{pmatrix}
q_{11} & q_{12} & \cdots & q_{1m-1} & q_{1m} \\
q_{21} & q_{22} & \cdots & q_{2m-1} & q_{2m} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
q_{n-1,1} & q_{n-1,2} & \cdots & q_{n-1,m-1} & q_{n-1,m} \\
q_{n,1} & q_{n,2} & \cdots & q_{n,m-1} & q_{n,m}
\end{pmatrix} \begin{pmatrix}
\phi^T(d_1) \\
\phi^T(d_2) \\
\vdots \\
\phi^T(d_{n-1}) \\
\phi^T(d_n)
\end{pmatrix}$$

The author uses the maximum co-occurrence strength between terms to add the semantic information in the traditional high-dimensional vector space model (only 0,1 weight), that is, if the weight value is 0 then give a new value between 0 and 1.

**Ternary CLSVSM**

In the literature [1], the author has proved that the CLSVSM based on the binary information can extract the latent semantic information, but only researching the binary information is not enough, and we need to extract the co-occurrence information beyond binary.

**The Representation of Ternary Co-occurrence Layer Matrix**

The primary question of ternary co-occurrence is how to represent the co-occurrence information. In this passage, we present a ternary co-occurrence layer matrix with algebraic method.
Definition 1: Given documents sets $D$ and its document-term matrix $A$, let $t_j$ be the j-th column of the matrix, we define co-occurrence of the j-th term $t_j$ and all term pairs $(t_i, t_k)$ $(i,k=1,2,\cdots,m)$ is the j-th layer of ternary co-occurrence as follow:

$$C_j^{(3)} = C_t^{(3)}(t_j) = (\text{diag}(t_j)A)^T \cdot (\text{diag}(t_j)A)$$

where $\text{diag}(t_j)$ is a diagonal matrix and its diagonal element is the component of $t_j$.

Firstly we calculate $\text{diag}(t_j)A$. In calculation, we use $\text{diag}(t_j)$ to modify the document-term matrix $A$ and constitute a new matrix, the modification method is as follow: If the diagonal elements of matrix $\text{diag}(t_j)$ contain 0, that is, the term $t_j$ at least in one document (the document set is $d$) does not appear, at this time the k-th row and the k-th column elements are 0. According to the matrix multiplication, the kth row the elements of the $\text{diag}(t_j)A$ are all 0. If the diagonal elements of $\text{diag}(t_j)$ are all 1, that is, the term $t_j$ appears in all documents, the result of $\text{diag}(t_j)A$ is $A$. Then we calculate the $(\text{diag}(t_j)A)^T \cdot (\text{diag}(t_j)A)$, and the form is similar to the binary co-occurrence matrix.

For example:

$$A = (t_1, t_2, t_3, t_4, t_5) = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

$$\text{diag}(t_1)A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

$$C_1^{(3)} = C^{(3)}(t_1) = (\text{diag}(t_1)A)^T \cdot (\text{diag}(t_1)A) = \begin{bmatrix} 4 & 3 & 2 & 2 & 3 \\ 3 & 3 & 2 & 2 & 2 \\ 2 & 2 & 2 & 1 & 1 \\ 2 & 2 & 1 & 2 & 1 \\ 3 & 2 & 1 & 1 & 3 \end{bmatrix}$$
In this calculation, we find that the diagonal elements of \( \text{diag}(t_j) \) contains 0, that is, the elements of the second row and the second column are all 0, namely the first term is not belong to the second document, therefore the second row of \( \text{diag}(t_j)A \) is 0. \( C_1^{(3)} \) is the co-occurrence frequency matrix of the first term \( t_1 \) with other terms. Obviously, the matrix is a symmetric matrix. The second row third column \( C_1^{(3)}(2,3) = 2 \) is the co-occurrence frequency of the terms \( t_1, t_2 \) and \( t_3 \). It is easy to verify that the matrix is correct, and the three words appear simultaneously in the first line and the fifth line.

The Relative Strength of Ternary Co-occurrence

It is not difficult to extract the ternary co-occurrence frequency through the program and even the frequency that more than ternary can be obtained. This paper mainly studies the calculation method of ternary co-occurrence frequency.

Firstly, we calculate the ternary co-occurrence frequency, let \( C_i, C_j, C_k \) respectively be the frequency of terms \( t_i, t_j, t_k \) and let \( C_{ij}, C_{jk}, C_{ik} \) be the frequency of the binary co-occurrence. \( C_{ijk} \) is the frequency of the ternary co-occurrence. Let \( c_{i\cup j\cup k} \) be the number of documents that including at least one of the three terms. Contrasting the promotion of the addition formula in the probability theory, for the three events A, B, C, let \( P \) be the probability of occurrence of events, then we define \( P(A\cup B\cup C) \) as follow:

\[
P(A\cup B\cup C) = P(A) + P(B) + P(C) - P(A\cap B) - P(A\cap C) - P(B\cap C) + P(A\cap B\cap C)
\]

We can infer:

\[
C_{i\cup j\cup k} = C_i + C_j + C_k - C_{ij} - C_{jk} - C_{ik} + C_{ijk}
\]

\[
C_{ijk} = C_{i\cup j\cup k} - (C_i + C_j + C_k) + (C_{ij} + C_{jk} + C_{ik})
\]

In this formula, \( C_{i\cup j\cup k} \) can be obtained from the document-term matrix, \( C_i, C_j, C_k, C_{ij}, C_{jk}, C_{ik} \) can be obtained by the co-occurrence matrix between terms, and then we can calculate the ternary co-occurrence frequencies.

Definition 2: given the set of documents \( D \) and its document-term matrix \( A \) and binary co-occurrence matrix \( C \), we can define the relative strength of ternary co-occurrence \( b_{ijk} \) of any three terms \( (t_i, t_j, t_k) \) as follow:

\[
b_{ijk} = \frac{C_{ijk}}{\sqrt[3]{|C_{ij}C_{jk}C_{ik}|}}, i, j, k = 1, 2, \ldots, m
\]

The Weighted CLSVSM Based on Binary and Ternary

In the new CLSVSM, we consider combining with binary and ternary co-occurrence information, and give the ternary greater weight.

Definition 3: Given the set of documents \( D \) and its document-term matrix \( A \), the binary co-occurrence relative strength matrix \( B \), the ternary co-occurrence matrix \( C^{(3)} \), the j-th floor
ternary co-occurrence matrix $C_{j}^{(3)}$, the j-th floor ternary co-occurrence relative strength matrix $B_{j}^{(3)} = (b_{ijk})_{m \times m}$. We define the new co-occurrence latent semantic vector space model based on binary and ternary as follow:

$$
\varphi(d_{i}) = \tilde{d}_{i} = (q_{i1}, q_{i2}, \ldots, q_{im}) \in \mathbb{R}^{m}
$$

$$
q_{ij} = \begin{cases}
1, & a_{ij} = 1 \\
\max_{rel_{i}} \{b_{ij}\}, & a_{ij} = 0, \max_{rel_{i}} \{b_{ij}\} \neq 0, \max_{rel_{i}} \{b_{ij}\} = 0 \\
\frac{1}{3} \max_{rel_{i}} \{b_{ij}\} + \frac{2}{3} \max_{rel_{i}, rel_{j}} \{b_{ij}\}, & a_{ij} = 0, \max_{rel_{i}} \{b_{ij}\} \neq 0, \max_{rel_{i}} \{b_{ij}\} = 0 \\
0, & \text{others}
\end{cases}
$$

Where, $I_{i} = \{ j | a_{ij} = 1 \}$ is the indicator sets of $j$.

The new “document - term” model is defined as:

$$
Q = \begin{pmatrix}
q_{11} & q_{12} & \cdots & q_{1m-1} & q_{1m} \\
q_{21} & q_{22} & \cdots & q_{2m-1} & q_{2m} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
q_{n-1,1} & q_{n-1,2} & \cdots & q_{n-1,m-1} & q_{n-1,m} \\
q_{n,1} & q_{n,2} & \cdots & q_{n,m-1} & q_{n,m}
\end{pmatrix}
= \begin{pmatrix}
\varphi^{T}(d_{1}) \\
\varphi^{T}(d_{2}) \\
\vdots \\
\varphi^{T}(d_{n-1}) \\
\varphi^{T}(d_{n})
\end{pmatrix}
$$

Experiment and result analysis

Experimental Data

In this experiment, we use the Chinese and English data to compare the binary CLSVSM and the weighted CLSVSM results respectively.

The Chinese data sets were collected in disciplines under three categories of information science in CNKI. After getting rid of non keyword documents, we get a total of 896 documents, which includes 299 "published", 298 "Library and Information and Digital Library", and 299 "archives and museums". The number of key words is 2509 (without repetition).

The English data sets used in the experiment are from "web of science". Firstly, we retrieve the first 1500 articles in the information science and library science according to the cited frequency in descending order, if the frequency is the same, we choose according to the system automatic sort. After getting rid of non keyword documents, we get a total of 411 documents, which includes 234 "Computer Science Information System", 123 "Management", and 54 "Computer Science Interdisciplinary Applications". The number of key words is 1661 (without repetition).

Experimental Process

This paper compares the two models according to the following steps:

Step 1: Pre-process the data and extract the terms of each document;
Step 2: Calculate the frequency of each term, list the terms in descending order and construct the co-occurrence matrix;
Step 3: Construct the vector space model and establish the document-term matrix of the literature;
Step 4: Calculate the co-word matrix according to the document-term matrix, get the frequency of the keywords, the binary co-occurrence, and the ternary co-occurrence;
Step 5: Calculate the relative strength of the binary and ternary co-occurrence, calculate the binary CLSVSM and the weighted CLSVSM, there will be two new document-term matrix;
Step 6: Use CLUTO software for clustering and it mainly uses the two new matrix;
Step 7: Record the experimental data, compare the experimental result, and come to the conclusion;

In this paper, 50 experiments were carried out on Chinese and English data respectively. The number of clusters was 3, and the cosine similarity was used. The number of iterations was 10 times for each experiment. From solution 1 to solution 50, we all use D-I2 program. That is, the clustering algorithm is Direct K-means, and the criterion function is I2. In each experiment, we have to record the value of purity and entropy, after the experiment, we also need to calculate the BF value. Finally, we also need to calculate the mean and standard deviation of purity, entropy, BF value and compare the results of the two models.

If the experimental subjects are divided into \( k \) classes, defined as \( L_j (1 \leq j \leq k) \). They can be clustered into \( k \) clusters, wrote as \( S_r (1 \leq r \leq k) \). The number of articles in the literature is \( n \), \( S_r \) and \( L_j \) contain \( n_r \) and \( n_j \) documents respectively. There are \( r_j \) identical documents.

Purity and entropy can be defined as follow:

\[
Purity = \sum_{r=1}^{k} \frac{1}{n} \max_{1 \leq j \leq k} n_{rj} = \frac{1}{n} \sum_{r=1}^{k} \max_{1 \leq j \leq k} n_{rj}
\]

\[
Entropy = \sum_{r=1}^{k} \frac{n_{rj}}{n} \left( -\frac{1}{\log k} \sum_{j=1}^{k} \frac{n_{rj}}{n_r} \log \frac{n_{rj}}{n_r} \right)
\]

B-Cubed method is the evaluation index of information retrieval, the main idea is to calculate the accuracy of each class (Precision) and recall (Recall), then calculate the weighted average sum and get the overall precision and recall rate. Precision and recall are defined as follows:

\[
BP_r = \text{precision}(r, j) = \frac{n_{rj}}{n_r}
\]

\[
BR_j = \text{recall}(r, j) = \frac{n_{rj}}{n_j}
\]

And the total precision and recall are:

\[
BP = \frac{1}{n} \sum_{r=1}^{k} n_r \times BP_r
\]

\[
BR = \frac{1}{n} \sum_{j=1}^{k} n_j \times BR_j
\]

BF value can be defined as follow:

\[
BF = \frac{2 \times BP \times BR}{BP + BR}
\]
Experimental Results

In this paper, the weighted CLSVSM experiment is compared with the binary experiment. The results of the Chinese data sets and the English data sets are shown in Table 4.1 and Table 4.2.

### Table 4.1 Comparison of two models (Chinese data)

<table>
<thead>
<tr>
<th></th>
<th>Entropy</th>
<th>Purity</th>
<th>BF Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted CLSVSM</td>
<td>0.579±0.029</td>
<td>0.784±0.016</td>
<td>0.791±0.013</td>
</tr>
<tr>
<td>Binary CLSVSM</td>
<td>0.596±0.039</td>
<td>0.768±0.037</td>
<td>0.776±0.034</td>
</tr>
</tbody>
</table>

Note: The data in the table are mean ± standard deviation

In the three clustering indicators, the closer the purity and BF values are to 1 and the entropy is to 0, the better the clustering effect. That is, the clustering effect will increase with the enlargement of purity and F value, and decrease with the enlargement of entropy. In the table 4.1, we first analyze the value of mean, it can be clearly seen that purity and BF value of the weighted model are higher than the binary experiment; the entropy is lower than the binary experiment. And then analyze the value of mean ± standard deviation, according to the basic knowledge of the statistics; we can see the value within the range of mean ± standard deviation, which is a large probability. We take the BF value as an example, in the weighted model, the value falls in [0.778, 0.804] is a high probability event, and the binary CLSVSM experiment is a large probability event in the range of [0.742, 0.8]. Relatively speaking, the value in the weighted model is bigger. In general, the clustering effect of the weighted model is better than binary model.

### Table 4.2 Comparison of two models (English)

<table>
<thead>
<tr>
<th></th>
<th>Entropy</th>
<th>Purity</th>
<th>BF Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted CLSVSM</td>
<td>0.552±0.011</td>
<td>0.729±0.004</td>
<td>0.681±0.015</td>
</tr>
<tr>
<td>Binary CLSVSM</td>
<td>0.545±0.004</td>
<td>0.718±0.005</td>
<td>0.682±0.012</td>
</tr>
</tbody>
</table>

In the table 4.2, the result of the weighted CLSVSM model is closer to the binary model, which is due to the diversity of the documents we choose. The possible reason is that there is a phenomenon of polysemy in English term, but we ignored it.

In the CLSVSM, the weighted CLSVSM experiment effect is better than the binary experiment in the Chinese literature. For the English literature, because the choice of literature and the processing of the terms are not perfect, the clustering results of the two models are similar in this experiment.

Conclusions

This paper mainly extended the binary co-occurrence model; we researched the representation of ternary co-occurrence, the calculation of ternary co-occurrence frequency and the relative strength, and formation of the weighted CLSVSM, and gave the experimental design. We showed that, in the Chinese literature, the clustering effect of the weighted model was better but in the English literature, the two clustering effects were equivalent. In this paper, the representation of the ternary co-occurrence layer matrix can be extended to the multiple co-occurrences, but it should be noted that in general the frequency of the multiple co-occurrence will decrease rapidly with the increase of multiple numbers. The multiple co-occurrence phrases are also rapidly decreasing. In the subsequent study, we will try to find a way to directly represent the ternary information and explore the scope of the model.

In addition, in the weighted CLSVSM model, we haven’t considered whether the weight value is the most appropriate value. And another question I want to say is that whether the
choice of terms in this paper (keywords) is enough. The terms extracted from references or the title of each document can also reveal the relations of the documents.

References
**Introduction**

Cancer, brain cancer as subtype, more threatening than of “AIDS, malaria and tuberculosis combined”, revealed and predicted its alarming signs to increase 80% in 2030. In concepts mapping for medical fields, various data mining techniques have been used, although mutual interaction and dynamics of those concepts are still unclear (Kostoff 2014). Such interational studies could track the behavioural dynamics of such concepts/factors to identify and understand the emerging technologies and drifts in concepts.

To face this challenge, we tried to develop a system, Hybrid Relation Analysis (HRA) for concept mapping in connections with relations, where we used topic modelling and variant of references. In Bibliometrics, Latent Dirichlet Allocation (Blei, Ng et al. 2003), is a classical choice regarding its efficacy and practical implication (Blei 2011) and has been utilized for various localized purposes. Bibliographic coupling is way of indirect relation and rated better performance than citations and co-citations (Boyack and Klavans 2010).

Our objective is to develop a system to identify and detect the drift of emerging technologies and understand the concepts of resultant innovation, where mechanism, extent, dominance and temporal trend of technological concepts and their interaction can be measured. We used technological literatures of brain cancer that can be valuable to comprehend the commercial perception of cause and treatment relationships as well as its commercial insight.

**Method**

We used Derwent Innovation Index, a precise, integrated, value added with originality, competitive and opportunity oriented patent database to get patents records for the disease brain cancer from years 1987 to 2016.

Latest machine learning, text mining, bibliometrics, data visualization and higher programing language-based approaches, are applied for the analysing the various databases. We based our experiment on optimizing conjugate gradient and Gibbs sampling. All the experiments have been performed on Intel Core i5 with 2.9 GHz and 16 GB main memory.

The search strategy, Huang (Huang, Schuehle et al. 2015) is applied with customization to ensure maximum recall and precision rates using patents classification codes relating to brain (G01N-033/574) and cancer (G01N-033/574).

After applying various pre and post data processing techniques, topic modeling, classical Latent Dirichlet Allocation method (Blei, Ng et al. 2003) is preferred for its reliability, multi-labeling and generative characteristics.

Later, second phase utilize the references, where for optimum number of factors divergence measure (Kullback and Leibler 1951) for any two matrices is:

HRA, as variants to LDA, for two hypothetical factors Zk and Zk’ having same citation link, CM from M documents that are independent and not governed by the document-theme relatedness utilizing the value of Θ, which is probability of document d for given factors k (Fig 2).

**Figure 1. Framework**

**Figure 2. Plate Notation Difference between LDA and HRA**
**Case study**

The method, yet generative, aimed at patent database due to flexibility in citation manipulation and commercial implication, although our experience and domain knowledge can affect result insight.

**Results**

In Fig 3, shows the relations (shared references) among eight concepts (ball). Distance between the factors is semantic relevance i.e., a closer is the neighbour, related is the concepts. Size of the ball reflect total number of shared citations. Position is comparative place, and control by all the mentioned factors, and besides three highlights can be drawn:

- For eight hypothetical factors, 56 possible combinations ($N^2-N$, where $N$ = No. of factors) are presented with 28 unique combinations for distinctive mutual impacts;
- “T-2: Thyroid Gland”, “T-6: Indirect Diseases” and “T-5: Antibody” stand as effective influential factors;
- Shared citation, depict a distinct picture for technological concepts, and dynamics combined with concepts relatedness.

Moreover, Fig.4 depicts the comparison between the total references and respective sharing for all concepts.

![Figure 3. Visualization of HRA in the technological field, Brain Cancer](image)

![Figure 4. A comparison of total and shared references for each concept.](image)

**Discussion and Conclusion**

Our HRA scheme used topic modelling and shared citations as concerted approach for mutual interaction, and dependency of technological topics to detect and understand emerging technologies, as described by Kostoff et al. (Kostoff 2014) that how bibliographic coupling and shared text could serve the optimistically better role than solo impacts. This innovative approach could improve topic coherence, interaction and depict supplementary insight to understand emerging technologies, where cluster linkage improve the clarity of map (Boyack and Klavans 2010) better than some other measures like citations and co-citations. As it is mentioned, bibliographic coupling originally is a powerful tool to map the coherent themes and documents representing similar research focus. Beside the total number of references are low, their sharing is dramatically higher in number for any concepts (Fig 4.).

All possible combinations follow: $N^2-N$; $N>1$, $N$= No. of factors. The study depicts that regulatory chemicals, secondary disorders and natural defence mechanism are the dominant and emerging concepts in field of brain cancer. Future research could focus on more specific implication of these concepts like which hormones, distinctive various diseases and components blood cells (immune system) and as major determinants.

**Acknowledgments**

The study supported by the National High Technology Research and Development Program of China (Grant No.2014AA015105) and Chinese Government Scholarship (2014-CSC 278401 ).

**References**


Boyack, K. W. and R. Klavans (2010). "Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately?" Journal of the American Society for Information Science and Technology 61(12): 2389-2404.


Dong Kun1  Xu haiyun 1  Luo Rui1  Yue Zenghui 2  Wei Ling 1

1 dongkun@mail.las.ac.cn  xuh@clas.ac.cn  15651006975@163.com  weiling@mail.las.ac.cn
Chengdu Documentation and Information Center, Chinese Academy of Sciences, Chengdu (China)
University of Chinese Academy of Sciences, Beijing (China)

2 yzh66123@126.com
School of Medical Information Engineering, Jining (China)

Abstract
Given that many frontiers and hotspots of science and technology are emerging from interdisciplines, the accurate identification and forecasting of interdisciplinary topics has become increasingly significant. Existing methods of interdisciplinary topic identification have their respective application fields, and each identification result can help researchers acquire partial characteristics of interdisciplinary topics. This paper offers an integrated method for identifying and predicting interdisciplinary topics from scientific literature. It integrates various methods, including co-occurrence networks analysis, high-TI terms analysis and burst detection, and offers an overall perspective into interdisciplinary topic identification. The results of the different methods are mutually confirmed and complemented, further overviewing the characteristics of the interdisciplinary field and highlighting the importance or potential of interdisciplinary topics. In this study, Information Science & Library Science is selected as a case study. The research has clearly shown that a more accurate and comprehensive results can be achieved for interdisciplinary topic identification and prediction by employing this integrated method. Further, the integration of different methods has promising potential for application in knowledge discovery and scientific measurement in the future.

Conference Topic
Research Fronts and Emerging Issues
Knowledge discovery and data mining
Methods and techniques
Co-occurrence analysis
Mapping and visualization
Social network analysis
Introduction
With the development of science and technology, scientific research is no longer limited to the study of a single field, but extended to interdisciplinary research. Bioinformatics, an interdisciplinary between Molecular Biology and Computer Science, is an excellent example to show the interdisciplinary trend because it utilizes computer tools to store, retrieve, and analyze biological information. In fact, new theories of science or emergence of inventions often originate from the intersections of disciplines, and the emphasis on interdisciplinary studies will lead to more precise and comprehensive results for frontiers prediction of science and technology.

Scientific literature is an important data source for interdisciplinary research. It is significant to explore effective knowledge mining algorithms to identify interdisciplinary topics from mass scientific literature and to make predictions. Interdisciplinary topics identification is different from conventional hotspots and frontier topics recognition, which not only reflects the popularity and growth of disciplinary knowledge, but also reflects its intersectionality. Traditional interdisciplinary analysis, which is based on high frequency terms, has certain one-sided limitations that make it bound to miss some important information and fail to guarantee the accuracy of the results. Therefore, developing an integrated method that makes full use of the characteristics of multiple topic identification methods is a pre-requisite.

This paper recalls various methods of topic identification, clarifies the characteristics of different methods in interdisciplinary topics analysis, and then puts forward an integrated method for interdisciplinary topic identification and prediction, offering an overall perspective into interdisciplinary research.

The paper continues with the presentation of the literature review. After the research methodology, the subsequent section presents the results and discussion. Finally, the paper concludes with some positive effects and limitations of the method.

Literature review
Many methods for topic identification have been provided in existing research, which mainly used knowledge units as the database. Knowledge units consist of terms, phrases, sentences, paragraphs and the like, among which topic terms analysis has the widest range of application, and types of topic terms cover both keywords given by authors and terms abstracted from the text by natural language processing (NLP).

High-frequency term analysis is the most basic and important method in topic identification, which reveals the research focus through frequency statistics. It is considered that terms of higher frequency are more likely to represent the hotspots in the field (Chu et al. 2014). By using high-frequency term analysis, many researchers took the relationship between terms into consideration. Co-occurrence analysis of terms is a typical method to meter the co-present situation of two terms that appeared in the same document, and then reflect the special relationship among them using cluster analysis (Cui 2003). It can reveal the link of research topics represented by topic terms, analyze the topic structure, and indicate the field’s hotspots and trends (Garfield 2004; Rokaya 2008; Xie 2016). However, the vast majority of co-occurrence analysis was merely based on strong co-occurrence networks. It means that terms have relative low co-occurrence frequency will be ignored, which leads to the exclusion of some potential and important topics, especially interdisciplinary topics. Wei et al. (2015)
constructed several weak co-occurrence networks to analyze research topics and their intersectionality. According to the research, the less highlighted topics in the strong co-occurrence networks would become obvious in the weak co-occurrence networks, which may be the current or potential focus.

Burst detection is another effective method for topic identification, which has a unique advantage in revealing the dynamic trends of terms and topics. It is known as burst term if the term suddenly increases in a short time, and generally, it does not belong to high frequency terms (Hong 2010). Several scholars have applied burst detection in research frontier and future trends analysis (Mane 2004; Wang 2015). Kleinberg (2003) constructed a burst detection algorithm to discover burst phenomenon, which has been the most widely used algorithm in burst detection based on scientific literature. The algorithm focuses on the terms whose relative growth rates suddenly increase, and it has been applied in multiple studies to predict the evolution trend of specific research area (Zhou 2015; Li 2017) due to its scientificty and convenience.

Many authors investigate interdisciplinary topic identification based on the above methods (Li 2013; Min 2014). Meanwhile, new attempts are made by employing some indexes. Xu (2015) used a series of indexes for interdisciplinary topic mining called high-TI terms analysis, which consists of two main indexes. The first index is TI value, which can be calculated depending on the formula $TI = d \times \log tf$; $d$ is the number of disciplines covered by the term, and $tf$ is the term frequency. This method took the distribution of terms into consideration, and reflected the interdisciplinary impact between disciplines. The other index is “betweenness centrality” (BET), a common social network analysis indicator. It was proposed by Freeman (1977) and was considered as the “intermediate” level of a node to the whole network. It can be obtained from some social network analysis software, such as UCINET. For either of the indexes, the interdisciplinarity has a positive association with the value. However, unlike TI value, which focuses on cross points between disciplines, BET value mainly represents the hinge points inside disciplines. These indexes can be efficient indicators for interdisciplinary analysis when complemented with other methods.

Above all, high-frequency term analysis and strong co-occurrence networks analysis based on high frequency terms have the widest range of applications because they generalize the most useful information. At the same time, the growing trend of using the terms and other subtle but valuable information may be ignored. Burst detection, which successfully focuses on the term’s dynamic tendency, has superiority in predicting future interdisciplinary topic and its results are highly likely to be hot topics. Weak co-occurrence networks analysis can only reflect partial and inconspicuous characteristics of the research topics in the field, but they are important supplements for strong co-occurrence topics and are more likely to represent the intersection of different disciplines as away from the field’s core topics. High-TI terms analysis can directly and intuitively reveal the interdisciplinarity among topic terms despite the requirement that it must be combined with the co-occurrence networks analysis.

Given that the above analysis methods have their respective fields of application, it is difficult to determine the best method, which reflects that each method has its emphasis. Similarly, any identification result can help researchers to analyze only partial characteristics of interdisciplinary topics. It is only when identifying interdisciplinary topics from different perspectives that we can comprehensively reveal its overall characteristics.

In summary, this study aims to put forward an integrated method for interdisciplinary topic identification and prediction, which integrates various results obtained from different methods,
including co-occurrence networks analysis, high-TI terms analysis, and burst detection, to
identify interdisciplinary topics from different perspectives and make interdisciplinary topic
identification and prediction more comprehensive and precise.

Research methodology
The framework of the integrated method for interdisciplinary topic identification and
prediction is shown in Fig. 1, which includes four parts.

Data collection
Considering that Information Science & Library Science (LIS) is a typical interdisciplinary
discipline, LIS was chosen as a case study to show the application of the integrated method.
SCI-EXPANDED, SSCI, CPCI-S, CCR-EXPANDED, and IC databases of the Web of Science (WOS) were selected as the data sources. The search strategy was “WC=Information Science & Library Science”, and the publication type was limited to “Article”. Data from 2007 to 2016 were retrieved on 2nd January, and 35,270 items were finally accessed.

Data processing
For massive scientific literature, mining and cleaning of the terms are time-consuming and laborious. Thomson Data Analyzer (TDA) is powerful in data extraction and cleaning (TDA 2016), which we employ to process data with the advice of experts. First, the dataset was prepared by extracting the titles, abstracts and references, and other paragraphs in the first part of the literature data. Then, meaningful topic terms or phrases were extracted until a certain threshold was reached through text mining. Next, the terms were cleaned using the following steps, including cleaning based on general thesauri, depth cleaning based on fuzzy matching, terms merging, and terms clustering based on principal component analysis. The final topic terms were cleaned, and the dataset of topic terms was constructed.

Interdisciplinary topic identification
Interdisciplinary topic identification is at the core of the integrated method, comprises three procedures. First, the top 300 high frequency terms were selected as basic data, and TDA was utilized to obtain the co-occurrence 300 × 300 matrix. Then, the matrix was placed into Ucinet, and GN algorithm was selected to achieve the following co-occurrence networks by clustering and visualization: (1) Strong co-occurrence networks of high-frequency terms: in this network, the term was the only node, and the threshold value of co-occurrence frequency was set as 10. (2) Co-occurrence networks of high-frequency terms and disciplines: in this network, there are two types of nodes—terms and disciplines. The threshold value of co-occurrence frequency was set as 15. Next, the TI and BET values of every term were calculated in topic terms dataset, where each term was ordered in turn, and two lists were obtained. Then, the TI and BET values of every term were compared with the two co-occurrence networks mentioned previously. Terms were marked as square in strong co-occurrence networks of high-frequency terms and disciplines when their TI values were greater than 10. The terms were marked as a star in both networks when they show high TI value and high Bet value at the same time and appear simultaneously on both networks. Finally, the co-occurrence matrix obtained in the first step was imported into Gephi to obtain weak co-occurrence networks. To make it clearer, co-occurrence frequency ranged from 3 to 10, and the node degree was not less than 5.

The information above shows that three kinds of interdisciplinary topics can be obtained. Then, combining all the results and subtracting the significant ones enables the identification of interdisciplinary topics in a more systematic and wholesome manner.

Interdisciplinary topic predication
The prediction of interdisciplinary topics has two steps—time series analysis and burst detection. Time series analysis is based on the identification results in the previous part. We could determine the orientation of future evolution by analyzing the topics’ changing situations along a given timeline. In the study, the topics acquired through upper analysis would be lined up and the trends can be found in this process. Burst detection utilizes Sci2,
which is a knowledge discovery tool developed by Katy Börner (2009), a library and information expert at Indiana University, Bloomington. It could detect frontier terms in the field using Kleinberg's burst monitoring algorithm. In this research, we set the detection interval from 2007 to 2016 and took the burst terms as the representation of related topics. More efficient and accurate prediction results can be achieved by combining the two steps.

**Results and Discussion**

Considering the large amount of data, this study chose 2007, 2009, 2011, 2013, and 2015 as the target years and made a detailed analysis. Meanwhile, only visualization results of 2007 were presented owing to space constraints.

**Interdisciplinary topic identification in 2007**

Figure 2 shows a strong co-occurrence network of high-frequency terms, and it is divided into 14 sub-networks, including complex networks, technology application practice, open access, journal evaluation, digital libraries, document services, bibliometrics, information management, information technology, information retrieval, information science, citation analysis, user study, and knowledge management.

![Fig. 2. Strong co-occurrence network of high-frequency terms in 2007.](image)

Figure 3 presents the co-occurrence network of high-frequency terms and disciplines, which indicates that LIS mainly interacts with the History of Social Sciences, Computer Science, and Management in 2007. The high-TI and high-Bet terms appearing in two networks are found in information retrieval, information management, bibliometric analysis, information technology, information systems and empirical investigation. This implies that these topic terms are the key nodes when LIS interacts with other disciplines, and the meanings reflected by these terms are regarded as important hinges and cross nodes in interdisciplinary areas in 2007.
Figure 4 illustrates a weak co-occurrence network, which is broken down into four sub-networks—communication technologies, information science, digital libraries, and bibliometric analysis. All of these topics are also contained in the results of strong co-occurrence network of high-frequency terms, but the positions of bibliometric analysis, communication technologies, and digital libraries in the weak co-occurrence network are more prominent in proportion to their importance, manifesting that these interdisciplinary topics have the potential for further development and should not be underestimated. Moreover, the bibliometric analysis is also an important hinge and cross node in high-TI terms analysis,
indicating that the importance of some topics has been highlighted by the weak co-occurrence networks analysis. In addition, the results of weak co-occurrence network analysis are consistent with the results of strong co-occurrence networks analysis and high-TI terms analysis to some extent, which proves that weak co-occurrence networks analysis is a necessary complement to the other two methods.

**Interdisciplinary topic identification in 2009**

The strong co-occurrence network of high-frequency terms in 2009 can be divided into 17 sub-networks, which include user research, e-government, complex networks, technology application practice, open access, digital libraries, text classification, bibliometrics, information services, information management, information technology, information retrieval, information systems, information literacy education/train, academic libraries, citation analysis, and knowledge management. LIS mainly interacts with Computer Science and Management. Important hinges and cross nodes are found in subjects including user satisfaction, open access, information management, citation data, information technology, information systems, complex networks, communication technologies, classification technique, bibliometric analysis, and semantic relationships analysis. The weak co-occurrence network is clustered into five sub-networks, which are information management, information technology, user satisfaction, open access, and scientific communication. They are all included in the strong co-occurrence network at the same time; however, the position of scientific communication in the weak co-occurrence network is more salient than in the strong co-occurrence network. Meanwhile, information management, information technology, user satisfaction, and open access are important hinges and cross nodes.

**Interdisciplinary topic identification in 2011**

The strong co-occurrence network of high-frequency terms in 2011 involves 14 sub-networks that contain web 2.0 technologies, technology study, competitive intelligence, open access, scientific collaboration, journal evaluation, information services, information behavior research, information technology, information retrieval, academic evaluation, academic libraries, citation analysis, and knowledge management. LIS interacts with Arts & Humanities, Computer Science and Business & Economics mainly by the terms, such as information technology, user acceptance, and information systems in 2011. The weak co-occurrence network in 2011 has six sub-networks, including information science, information retrieval, social networks, information technology, digital libraries, and knowledge sharing. All of the sub-networks are included in the strong co-occurrence network except the digital libraries. In addition, the positions of nodes of information science, information retrieval, social networks, and digital libraries in the weak co-occurrence network are more important than in the strong co-occurrence network. “Digital libraries” is a significant interdisciplinary topic with great potential, to which more attention should be given. Meanwhile, information technology is an important hinge and cross node from the perspective of high-TI terms analysis.

**Interdisciplinary topic identification in 2013**

There are 15 sub-networks in the strong co-occurrence network of high-frequency terms of 2013, including competitive intelligence, open access, scientific collaboration, journal evaluation, social network research, web use, bibliometrics, information behavior research,
academic evaluation, academic libraries, user study, semantic analysis, knowledge management, public libraries, and information sources. Although information technology, information systems, digital libraries, information retrieval, information science, and social networks were not classified into any sub-networks, they are important hinge nodes in the network. LIS usually interacts with Computer Science, Business & Economics, Health Care Sciences & Services, and Medical Informatics using the terms, such as user acceptance, information technology, information retrieval, information systems, social networks, knowledge sharing, and social media, which are important hinges and cross nodes. The weak co-occurrence network in 2013 is clustered into five sub-networks, which consists of information science, knowledge management, user acceptance, citation analysis, and information need. These topics are all contained in the strong co-occurrence network as well. The difference lies in the fact that the positions of knowledge management, citation analysis, and information need are more conspicuous than in the strong co-occurrence network. In addition, user acceptance is the hinge and cross node in high-TI terms analysis.

Interdisciplinary topics identification in 2015
The strong co-occurrence network of high-frequency terms of 2015 is divided into 11 sub-networks covering academic libraries, technology application practice, information behavior research, knowledge sharing, information science, complex networks, bibliometric analysis, competitive intelligence, information technology, text analysis, and digital libraries. LIS mainly interacts with Computer Science and Management. At the same time, it has interaction with Health Care Sciences & Services, Medical Informatics, and Communication to some extent. The important hinges and cross nodes are knowledge sharing, information management, qualitative analysis, information systems, academic evaluation, citation analysis, information science, complex networks, information technology, social media, and health informatics. It implies that the interdisciplinary features of these terms are prominent. The weak co-occurrence network of 2015 is clustered into five sub-networks covering qualitative analysis, social media, information systems, information management, and citation analysis, which all belong to the strong co-occurrence network. The positions of qualitative analysis, social media, information systems, and information management are more obvious than in the strong co-occurrence network. Meanwhile, the core nodes of the five sub-networks are all important hinges and cross nodes according to the results of high-TI terms analysis.

Interdisciplinary topic prediction
(1) Time series analysis
The results of time series analysis can be concluded as follows: First, some topics, which have been around in the co-currency networks for more than four years, may be considered as the most constant and hot interdisciplinary topics in the area. These topics cover open access or knowledge sharing or scientific communication, information management or knowledge management, information technology or communication technology, information retrieval, citation analysis, information system, technical application practice, journal evaluation or academic evaluation, bibliometric analysis, information science, user study or user satisfaction or user needs or user acceptance. Second, a few topics have disappeared in recent years, signaling that the importance of these topics has somewhat reduced and probably will continue to show a downward trend in the future. These topics cover literature service or
information service, e-government, text classification or classification techniques, information literacy education, scientific communication and others. Third, many topics have always existed since it appeared in the last few years. Undoubtedly, this group reflects the major orientation of interdisciplinary evolution in the near future. Semantic relationships analysis, competitive advantage, information behavior, scientific collaboration, social networks, and social media all belong to this group.

(2) Burst detection

The original results of Sci2 consisted of 323 terms, which were further screened out by artificial interpretation based on the correlation with research objects, and then 55 terms were filtered. Table 1 lists the burst terms that appear from 2014 and still persist. “Weight” represents the degree of emergency. “Start” represents the start year when a term began to be a burst term, and in a similar way, “End” represents the end year when a term is no longer a burst term. “Length” is the time span from start to end.

<table>
<thead>
<tr>
<th>Burst words</th>
<th>Weight</th>
<th>Length</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social media</td>
<td>26.99918</td>
<td>4</td>
<td>2014</td>
<td>2017</td>
</tr>
<tr>
<td>Information overload</td>
<td>3.268151</td>
<td>4</td>
<td>2014</td>
<td>2017</td>
</tr>
<tr>
<td>Organizational-behavior</td>
<td>3.038309</td>
<td>4</td>
<td>2014</td>
<td>2017</td>
</tr>
<tr>
<td>Societal impact</td>
<td>3.038309</td>
<td>4</td>
<td>2014</td>
<td>2017</td>
</tr>
<tr>
<td>Collaboration networks</td>
<td>2.711755</td>
<td>4</td>
<td>2014</td>
<td>2017</td>
</tr>
<tr>
<td>Publication patterns</td>
<td>2.574163</td>
<td>4</td>
<td>2014</td>
<td>2017</td>
</tr>
<tr>
<td>Altmetrics</td>
<td>10.89324</td>
<td>3</td>
<td>2015</td>
<td>2017</td>
</tr>
<tr>
<td>Virtual communities</td>
<td>9.672103</td>
<td>3</td>
<td>2015</td>
<td>2017</td>
</tr>
<tr>
<td>Big data</td>
<td>8.977111</td>
<td>3</td>
<td>2015</td>
<td>2017</td>
</tr>
<tr>
<td>Social network</td>
<td>7.829662</td>
<td>3</td>
<td>2015</td>
<td>2017</td>
</tr>
<tr>
<td>Web 2.0</td>
<td>6.544265</td>
<td>3</td>
<td>2015</td>
<td>2017</td>
</tr>
<tr>
<td>Co-authorship networks</td>
<td>5.020775</td>
<td>3</td>
<td>2015</td>
<td>2017</td>
</tr>
<tr>
<td>Scientific collaboration</td>
<td>6.715204</td>
<td>2</td>
<td>2016</td>
<td>2017</td>
</tr>
<tr>
<td>Social networking sites</td>
<td>4.972712</td>
<td>2</td>
<td>2016</td>
<td>2017</td>
</tr>
<tr>
<td>Regression-analysis</td>
<td>3.038755</td>
<td>2</td>
<td>2016</td>
<td>2017</td>
</tr>
<tr>
<td>Health literacy</td>
<td>2.986586</td>
<td>2</td>
<td>2016</td>
<td>2017</td>
</tr>
<tr>
<td>Page rank</td>
<td>2.746356</td>
<td>2</td>
<td>2016</td>
<td>2017</td>
</tr>
</tbody>
</table>

Given that the end year of all terms in the table is 2017, it can indicate that any term should not be underestimated in the following years. From the view of weight, we can see that social media has the highest value, revealing its greater significance in future studies. Next is altmetrics, a new method of assessing the impact of published works. With the continuous development of social media, altmetrics will be widely used in the research. Besides, the weight of virtual communities, big data, social network, scientific collaboration, web 2.0, co-authorship networks all exceed five, demonstrating the good development prospect of these topics. From the perspective of length, terms have a higher value, possessing both abruptness and a certain continuity. These terms include social media, information overload,
organizational-behavior, societal impact, collaboration networks, and publication patterns, whose Length values are greater than four.

Discussion
Based on the above analysis, the key characteristics of the interdisciplinary situation in LIS can be outlined. First, the number of interdisciplines connected to LIS shows an increasing trend. Among them, Computer Science is an enduring discipline, which shows the greatest similarity in research content to LIS, and Management and Business & Economics come second in light of similarity in research content. Medical Informatics and Health Care Sciences & Services demonstrate an obvious trend to interact with LIS. Second, the most important interdisciplinary topics in LIS are not only core nodes of strong co-occurrence networks of high-frequency terms, but also sub-networks of weak co-occurrence network of high-frequency terms, as well as terms with high TI and high BET value. These topics include bibliometrics (2007), information management, information technology, user satisfaction, and open access (2009), information technology (2011), user acceptance (2013), qualitative analysis, social media, information systems, information management, and citation analysis (2015). Third, the major future interdisciplinary topics consist of semantic relationships analysis, competitive advantage, information behavior, scientific collaboration, social networks, social media, altmetrics, virtual communities, big data, web 2.0, and co-authorship networks, among which social media, social network and scientific collaboration are especially on the rise due to the continuous development of communication tools.

Conclusion
In this study, we constructed an integrated method for interdisciplinary topic identification and prediction. By employing the integrated method, more accurate and comprehensive results can be obtained for interdisciplinary topic identification and prediction. The results of the different methods are mutually confirmed and complemented, further overviewing the characteristics of the interdisciplinary field and highlighting the importance or potential of interdisciplinary topics.

However, this integrated method is limited to the fusion of analysis results, which is only the late fusion stage for integration. Beyond that, early fusion stage (data integration) and mid-fusion stage (fusion of data relations) are also available and valuable. Thus, in the future, we plan to consider deep integration from the early or middle stage. Moreover, it is necessary to expand the integrated method further by absorbing other good methods into its system and exploring more possibilities of integration.

The integration of different methods has promising potential for application in knowledge discovery and scientific measurement in the future. However, the kind of objects that can be integrated, the choice between data types or data relation, and their method of integration, which is either artificial or automatic, require us to conduct more in-depth and long-term exploration.

Acknowledgments
This study was supported by the China Postdoctoral Science Foundation funded project (2016M590124), the Youth Innovation Fund of Promotion Association, CAS (2016159) and Intellectual Property Research & Consultation Center (IPRCC).
References


Prediction Model of Future Academic Impact for Young Scientists

Wei Tongqi\textsuperscript{1}, Lu Jingjing\textsuperscript{2}, Tan Zongying\textsuperscript{3}

\textsuperscript{1}weitq@psych.ac.cn
Institute of Psychology, Chinese Academy of Science, Beijing (China)

\textsuperscript{2}delialjj@126.com
People’s Public Security University of China, Beijing (China)

\textsuperscript{3}tanzy@mail.las.ac.cn
National Science Library, Chinese Academy of Science, Beijing (China)

Abstract
Previous studies focused on the characteristics of elite scientists, or the early identification of high-potential scientists, but neglected career development research of ordinary young scientists. Using scientometrics and CV methods, the present study aims to establish statistic models in order to explore young scientists’ career development, and predict their future academic impact. Predictors at 2009 were divided into four categories: publication indicators, scholar relationship indicators, scholar relationship indicators and research theme indicators, and these indicators were used to predict $h$-index at 2014. Predictive models were established separately in each category using stepwise regression, and finally a comprehensive model was built. Our result shows that $h$-index and total number of paper published could significantly predict scientific impact in the future. To improve academic impact, young scientists need to carry out extensive cooperation and play a key role in the cooperation, need to have the research experience in developed countries, and to reduce their research specialization.

Conference Topic
Studies on the level of individual scientists

Introduction
Whether we can base on small number indicators to predicate a scientists’ future scientific impact? More practically, for making a wise choice, which part of applicants’ CV should be pay more attention to, when scientific institutes want to recruit a new researcher? Many researchers had tried to answer these questions. Using machine-learning approach, Acuna et al. (2012) collected 3085 neuroscientists’ CV data, and match them with publication data in Scopus. They tried to use scientometrics indicators to predict $h$-index of these neuroscientists in 1 year, 5 year, and 10 years later. In a better linear predictive model, the predictors are current $h$-index, number of article written, years since publishing first article, number of distinct journals published in, and number of articles in Nature, Science, Nature Neuroscience, PNAS and Neuron. These indicators can explain 67% variance of $h$-index in the 5-years later. The work of Dijk et al. (2014) is also valuable. They tried to predict who will become a PI in the future based on 25000 researcher’s publication information in PubMed. The conclusion was that number of publications, the impact factor (IF) of the journals in which those papers are published, and the number of the papers that receive more citations than average for the journal in which they were published (citation/IF) are three important predictors. Whether a scientist will become PI depends to a great extent on their publication, especially in their early career. More importantly, this study found some non-publication features, such as the range of university and gender, also significantly determine whether a scientist will become a PI. Scientists’ CVs provide more information beyond their publication. Generally, CV contains many information about educational experience, mobility experience, gender, occupation years, scientific cooperation, research theme, and so on. Beside these scientometrics indicators calculated from their publication, whether these non-publication features obtained from CV --
- personal characteristics, scholar cooperation relationship, research theme --- could play a statistically significant role in predict future scientific impact? The current study tries to consider both of publication features and non-publication features, and finds out which indicators are more sensitive in predicting the future scientific impact after parameter comparison.

The literature review revealed that mobility experience is benefit scientific career development (Filippo, Casado, Gómez, 2009; Yamashita & Yoshinaga, 2014; Cañibano, Otamendi, Andúja, 2008; Jonkers & Tijssen, 2008). And other studies supported the opinion that post-doc experience enriches the researchers’ human and social capital, and has a positive effect on future academic performance (Zubieta, 2009; Gaughan & Robin, 2004). So, personal characteristic features in current study contain some scientific mobility indicators and post-doc indicators. Furthermore, because the differences on gender and career age must make certain influence on academic achievement (Cole, 1984; Symonds, Gemmell, Braisher, et al., 2011; Bornmann, Mutz, & Daniel, 2007; Dickersin, Fredman, Flegal, et al., 1998). Gender and age were also included into personal characteristic features.

Many studies also provided a lot of evidences that the scientific cooperation is positively related with scientific productivity and impact (Price, 1963; Zuckerman, 1967; Tang & Shapiro, 2012; Katz & Hicks, 1997). Recent researches used social network method to qualify one researcher’s co-author network (Abbasi, Chung, & Hossain, 2012; Mccarty, Jawitz, Hopkins, et al., 2013). In present study, we will also consider the size of co-author network and position in the network as the scholar cooperation feature.

Meanwhile, each researcher has their own special research field. Researchers in different research field would be have different career development trajectories. Whether the researcher who focus on hot topic or rising topic has more developmental potential? Whether one researcher consistently focused on one research field will acquire better scientific impact in the future than those who are interested in different fields? These research theme related features involve in theme hotness, theme novelty and researcher’s specialization.

In present study, different types of indicators will be included in a comprehensive predictive model to build an optimizing model with predictors as little as possible, and try to figure out the most effective factors. We also pay main attention to young scientists who just start their career path.

**Methods**

**Sample:** This study will focus on young psychological researchers of 15 psychological research institutes from USA, Netherland and China. The sample procedure excluded adjunct professor, emeritus professor, affiliate professor, doctor students, post-doctor and technical staffs. Only the scientists who got their doctoral degree between 2004 and 2009 were selected. Their CVs were downloaded from their websites and LinkedIn, and decoded. Their publication data was downloaded from Web of Sciences (WOS). And according to these data, we counted each person’s $h$-index in 2009 and 2014. The period of data collection was Mar – May, 2015.

**Indicators:** The dependent factor is $h$-index at 2014, as the measurement of future academic impact. The predictive variables were divided into four categories: publication indicator, scholar experience indicator, scholar relationships indicator, and research theme indicator. Eight publication indicators calculated from publication data were selected. They are $h$-index, total number of papers published, total citation, citation per paper, $h$-index per year, paper number per year, number of journals, and average impact factor. Seven scholar experience indicators decoded from CV are gender, whether has post-doc experience, duration and times of post-doc, career age, mobility frequency, and whether has work experience at American
institutes. Two scholar relationship indicators are total number of co-author and betweenness centrality of co-author network. We used each young psychologist as core node to build ego network, and calculated the betweenness centrality of core node. The authors who has high betweenness centrality are more important for their co-author network, and have more control on cooperation relationship. Research theme indicators are theme hotness, theme novelty and specialization. Theme hotness of each author was defined as the average number of $h$-index of all author keywords. Theme novelty of each author was defined as the average number of time span of all author keywords since their first appearance in publication. Leahey (2006) define the Specialization as “a researcher focusing tightly on one or a few subfields, rather than spanning many”. Calculation of specialization refers to the formula of Porter, et al. (2007). Specialization ($S$):

$$S = \frac{\sum[(\text{# of papers published in } WC_1)^2 + \cdots + (\text{# of papers published in } WC_n)^2]}{\sum[(\text{# of papers published in } WC_1 + \cdots + (\text{# of papers published in } WC_n)]^2}$$

Where WC is the Web of Science Category in WOS.

All the predictors used in present study see Table 1.

### Table 1. The predictors of the present study

<table>
<thead>
<tr>
<th>Category</th>
<th>Variables</th>
<th>Description</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publication</td>
<td>$h$-index</td>
<td>Number of papers with citation number &gt; $h$</td>
<td>H09</td>
</tr>
<tr>
<td></td>
<td>total number of papers</td>
<td>All journal publication in SCIE and SSCI</td>
<td>Num.JP09</td>
</tr>
<tr>
<td></td>
<td>published</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>total citation</td>
<td>Total citation in WOS</td>
<td>Citation09</td>
</tr>
<tr>
<td></td>
<td>citation per paper</td>
<td>Total citation divided by Num.JP09</td>
<td>AC09</td>
</tr>
<tr>
<td></td>
<td>$h$-index per year</td>
<td>$h$-index divided by Age.Career</td>
<td>AH09</td>
</tr>
<tr>
<td></td>
<td>paper number per year</td>
<td>Num.JP09 divided by Age.Career</td>
<td>AP09</td>
</tr>
<tr>
<td></td>
<td>number of journal</td>
<td>Number of journal published</td>
<td>Num.J</td>
</tr>
<tr>
<td></td>
<td>average impact factor</td>
<td>Total journal impact factor divided by Num.J</td>
<td>AIF</td>
</tr>
<tr>
<td>Scholar Experience</td>
<td>gender</td>
<td>Man, Woman</td>
<td>Gender</td>
</tr>
<tr>
<td></td>
<td>whether has post-doc</td>
<td>Since doctor gradated until 2009 whether has post-doc experience</td>
<td>PostDoc</td>
</tr>
<tr>
<td></td>
<td>experience</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Frequency of post-doc</td>
<td>Until 2009 the times of post-doc</td>
<td>Num.PostDoc</td>
</tr>
<tr>
<td></td>
<td>Duration of post-doc</td>
<td>all career years in post-doc position</td>
<td>Dur.PostDoc</td>
</tr>
<tr>
<td></td>
<td>Career age</td>
<td>Since start PhD research until 2009</td>
<td>Age.Career</td>
</tr>
<tr>
<td></td>
<td>times of mobility</td>
<td>Times of changing research institute or university</td>
<td>Num.Mobility</td>
</tr>
<tr>
<td></td>
<td>whether the destinations</td>
<td>Yes, No</td>
<td>Destination</td>
</tr>
<tr>
<td></td>
<td>include the American</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>institutes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scholar Relationship</td>
<td>network size</td>
<td>Number of co-author</td>
<td>Num.Co</td>
</tr>
</tbody>
</table>

1643
betweenness centrality of co-author network  |  Betweenness centrality of core node of ego-central network  |  Betweenness centrality of ego-central network  
---|---|---  
Research Theme  |  theme hotness  |  average number of \( h \)-index of all author keywords  |  Hotness  
  |  theme novelty  |  average number of time span of all author keywords since their first appearance  |  Novelty  
  |  specialization  |  focus tightly on one or a few subfields  |  Specialization  

**Procedure:** The retrospective analysis was used. The present study try to use the predictive variables in 2009 to predict academic impact of a specific scientist in 2014. Indicators in each category were used to find the most significant predictor in each model. Through stepwise regression, all predicative indicators were entered into formula one by one. According to Akaike Information Criterion, the non-significant variables were deleted, and the significant variables remained. The best fit model should satisfy the follow two conditions: all the variables in predictive model are significant, and number of predictors should keep as less as possible. Finally, using the same method, a comprehensive predictive model was built. R language were used to complete statistical analysis.

**Results and discussion**

**The predictive model of publication indicators:** Only two publication indicators can significantly predict the \( h \)-index in five years later. They are current \( h \)-index and number of publication. The R-squared for the obtained regression model is 69.83% (\( F=94.75, p<0.01 \)). There is no collinearity in this model. Variance Inflation Factor (VIF) is 3.82, below 10. The regression equation can be written as:

\[
H = 0.87 \, (H09) + 0.28 \, (Num.JP09) + 3.78
\]

This result is consistent with previous research conclusion that publication indicators could well predict academic impact in five years later (Acuna et al., 2012). What’s more, our result reveals that the model is also suitable for young scientists, even they just start their career development. Peterson, Jung, Yang, et al. (2011) developed a career progress model, and tested it on career of scientists and athletes. They found that cumulative advantage gained from past success will determine the future career development, an evidence of age-old Matthew “rich get richer” effect. Our model shows that for scientists, the Matthew effect appears very early. The tiny difference in \( h \)-index and number of publication between PhD graduates maybe became formidable lap in the later. Just as Peterson, Jung, Yang et al.’s study (2011), we also highlight the importance of early career development. Career competition started very early for scientists.

**The predictive model of scholar experience indicators:** In all seven variables, only the destination of mobility can significantly predict the \( h \) index in five years later. It means that the future scientific impact is not related with gender, post-doc experience, whether and number of scientific mobility, but the scholar experience in American institute seemly is important for future development, at least for psychologists. Meanwhile, its contribution is lower that current \( h \)-index and number of publication. The R-squared for the obtained regression model is only 16.32% (\( F=12.9, p<0.01 \)). The regression equation can be written as:

\[
H = 4.46 \, (Dest) + 7.26
\]
The predictive model of scholar relationship indicators: Both size and centrality of co-author network significantly predict h-index in five years later. The R-squared for the obtained regression model is 50.3% (F=40.98, p<0.01). There is no collinearity in this model. The regression equation can be written as:

\[ H = 0.19 \text{(Num.Co)} + 11.41 \text{(Betweenness)} + 4.41 \]

The size of co-author network reflects the wide cooperation relationship of scientist. McCarty, Jawitz, Hopkins, et al. (2013) found that the differences of h-index could be explained by number of co-authors. The current study confirms their conclusion, and further suggests that number of co-authors also can highly explain the variance of h-index in five years later. Moreover, the significant predictive effect of betweenness means that scientist should maintain the core position of their scholar network. The one who exhibits more control on cooperation network will have better career development in the future.

The predictive model of research theme indicators: Among three theme indicators, only specialization is the significant predictor of h-index in five years later. The contribution of specialization is 15.3% (F = 16.17, p< 0.01). The regression equation can be written as:

\[ H = -11.93 \text{(Specialization)} + 13.49 \]

The standard B value is negative. It means that if a scientist only focus on one research field, his/her scientific impact may become lower than his/her peers. This result seemingly violates common sense. But in practice, the scientist who only focus on certain scientific question indeed is difficult to publishing more research papers, and generally has small cooperation network.

The comprehensive full model: Based on four separate models, six significant variables were used to construct a multiple predictive model. Using the same lineal regression method, the comprehensive model was constructed. Three factors were eliminated, and other factors remain. H-index, number of publication and specialization still significantly predict h-index in the five years later. The final model is:

\[ H = 4.76 + 0.92 \text{(H09)} + 0.24 \text{(Num.JP09)} – 3.00 \text{(Specialization)} \]

The R-square of the model is 72.55% (F= 74.14, p<0.01).

We should not make causal interpretation of the relation between these predictors and future scientific impact. But empirically, just as the model shows, the young scientist who has higher h-index, published more papers, and has extensive research interest will have more potential in the future.

Conclusion
A empirical model was constructed. The model can use only three indictors --- h-index, total number of paper published, and specialization – to predict future research impact effectively. And the data also suggests early cumulative advantage is essential for future career development of young scientists. To improve academic impact, young scientists need to carry out extensive cooperation and play a key role in the cooperation, need to have the research experience in developed countries, and to reduce their research specialization.

References


Is the Gap in Scientific and Technological Strength Between G7 and BRICS Becoming Smaller or Larger?

Yuan Junpeng  Gao Jiping  Su Cheng  Zhai Lihua  Wang Haiyan  Pan Yuntao

junpengyu@gmail.com
Institute of Scientific and Technical Information of China, Beijing (China)

Introduction
Science and technology in emerging economies is undergoing rapid development. As a result, the question arises of whether the gap in scientific and technological strength between emerging economies and developed countries is becoming smaller or larger. Although there are many ways of studying scientific activity systematically, scientometric analysis provides a well-proven means of achieving this goal (Garfield E. 1979). Although numerous lists of highly cited articles have been compiled across a variety of scientific fields, few have included articles from all fields (van Raan 2017, Siddiqi and Usman 2017, Terekhov 2017). The present study was an in-depth scientometric analysis of data. It aimed to identify and analyse the 1000 most highly cited articles and evaluate the research performance of the G7 and BRICS countries. It is hoped that this study will stimulate useful discussion among scientists and research managers about research directions and provide insight for policy and funding decisions.

Dataset and methods
Since the reformation and opening up of its economy in 1978, China has undergone rapid development. The Web of Science (WoS) database was used to compile a citation history of papers for the period 1978–2008. Based on these articles and citations, the top 1000 highest-impact articles in every discipline were selected as the dataset. To measure scientific output, we adopted as indicators the total number of articles, total citation rates, citations per article, and numbers of domestic and international articles. To examine collaboration, we chose the index of international articles, the international cooperation rate to assess the role of international collaboration in high-impact articles.

Analyses and results
We consider whether a country’s stage of economic development and its S&T level can affect the production of its highly cited papers. Do S&T levels lead to or result from numbers of highly cited papers?

Characteristics of G7 countries’ highly cited papers
Domestic articles are taken to be highly cited papers that were written entirely by authors of one particular country. Analysis of the G7's domestic papers in 22 disciplines can be used as an indicator of the S&T capability and impact of the G7 countries in various disciplines. It can be seen that domestic papers of the G7 accounted for the majority of the papers in all the disciplines: except for the fields of space science, clinical medicine, agricultural sciences, and mathematics, those proportions were greater than 70%. The proportion of G7 domestic papers to all domestic papers in the discipline of general social sciences was more than 90%.

Because it provides access to a wider range of facilities and resources, collaboration is encouraged at a policy level. It enables researchers to participate in networks of cutting-edge and innovative activity. The international cooperation rate can reflect the G7 countries’ status of participation in international scientific and technological cooperation. The international cooperation rate of G7 countries in the 22 disciplines is shown in Figure 1.

Fig 1a. The international cooperation rate among the G7 countries in 22 disciplines

Fig 1b. The international cooperation rate among the G7 countries in 22 disciplines

The present study divided the years between 1979 and 2008 into three decades (1979–88, 1989–98, and 1999–2008) to compare the number of domestic articles in G7 countries in 22 disciplines. Analysis showed that the number of highly cited papers from 1979 to 1988 and from 1989 to 1998 was far greater
than the number of such papers from 1999 to 2008. This is in keeping with the principle that in the G7 countries papers published earlier tend to be cited more often.

**Characteristics of BRICS countries’ highly cited papers**

In terms of the number of papers in the various disciplines, there was a large gap between the rankings of the five BRICS countries and the G7 countries. Table 1 lists the total number of articles and the ranking by discipline of the BRICS countries. Table 1. The total number of articles and ranking by discipline of the BRICS countries (1979–2008)

<table>
<thead>
<tr>
<th>discipline</th>
<th>South Africa</th>
<th>China</th>
<th>India</th>
<th>Brazil</th>
<th>Russia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Materials Science</td>
<td>0</td>
<td>10(7)</td>
<td>3(15)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Geosciences</td>
<td>0</td>
<td>0</td>
<td>1(14)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Molecular Biology &amp; Genetics</td>
<td>0</td>
<td>0</td>
<td>1(13)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Engineering</td>
<td>1(24)</td>
<td>2(17)</td>
<td>3(15)</td>
<td>1(24)</td>
<td>1(24)</td>
</tr>
<tr>
<td>Chemistry</td>
<td>1(21)</td>
<td>0</td>
<td>2(16)</td>
<td>1(21)</td>
<td>1(21)</td>
</tr>
<tr>
<td>Environment/Ecology</td>
<td>2(17)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Computer Science</td>
<td>0</td>
<td>2(18)</td>
<td>3(15)</td>
<td>1(20)</td>
<td>0</td>
</tr>
<tr>
<td>Economics &amp; Business</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Psychiatry/Psychology</td>
<td>1(14)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Space Science</td>
<td>1(15)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Clinical Medicine</td>
<td>0</td>
<td>1(18)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immunology</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Agricultural Sciences</td>
<td>7(16)</td>
<td>1(26)</td>
<td>2(22)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Social Sciences, General</td>
<td>0</td>
<td>0</td>
<td>1(12)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Neuroscience &amp; Behavior</td>
<td>0</td>
<td>1(13)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Biology &amp; Biochemistry</td>
<td>1(19)</td>
<td>0</td>
<td>2(15)</td>
<td>0</td>
<td>1(19)</td>
</tr>
<tr>
<td>Mathematics</td>
<td>1(25)</td>
<td>3(19)</td>
<td>0</td>
<td>2(21)</td>
<td>2(21)</td>
</tr>
<tr>
<td>Microbiology</td>
<td>1(17)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Physics</td>
<td>1(18)</td>
<td>0</td>
<td>1(18)</td>
<td>1(14)</td>
<td>0</td>
</tr>
<tr>
<td>Pharmacology &amp; Toxicology</td>
<td>2(16)</td>
<td>0</td>
<td>0</td>
<td>1(19)</td>
<td>0</td>
</tr>
<tr>
<td>Plant &amp; Animal Science</td>
<td>4(15)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Multidisciplinary</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The economic strength and S&T capacity of the BRICS countries made substantial progress over the period from 1979 to 2008, and the number of SCI papers of the five countries also increased rapidly. It is evident that the highly cited papers in the BRICS countries were mainly the result of international cooperation. In terms of the average international cooperation rate among the BRICS countries, chemistry was the discipline with the lowest rate; the average rate amounted only to 16.67%. The next lowest were pharmacology & toxicology and mathematics. The other disciplines showed an average rate of greater than 50%. Notable was the rate of 100% in economics & business, immunology, and multidisciplinary, which indicates that the highly cited papers produced by BRICS countries in these three disciplines were entirely the result of international cooperation. Compared with the figures for the G7 countries, the average rate for chemistry and pharmacology & toxicology was also low among the BRICS countries.

**Conclusion and discussion**

The present study was designed to analyze the published articles that attracted the most citations in the G7 and BRICS countries. Some valuable insights can be gained through this restricted approach.

1) The United States was the world leader, followed by the United Kingdom, Germany, and Japan. There was a significant gap between Italy and the other G7 countries.

2) In terms of the number of papers in the various disciplines, there was a large gap between the rankings of the BRICS countries and the G7 countries.

3) There was a world trend toward domestic research in terms of highly cited papers. The papers in the majority of disciplines in the United States were based on independent domestic research. The United Kingdom, France, Germany, and Italy placed greater emphasis on international cooperation, and Japan and Canada had different tendencies toward international cooperation in different disciplines.

4) The highly cited papers among BRICS countries in economics & business, immunology, and multidisciplinary were produced entirely by international cooperation.

5) The growth rates were high in 13 disciplines in Germany and 12 disciplines in Japan. In the United Kingdom and France, the growth rates were high in eight disciplines. Finally, the United States had high growth rates in five disciplines, and in Canada and Italy growth rates were high in four disciplines.

**Acknowledgments**

This paper was supported by a grant (No: 71473236) from the National Natural Science Foundation of China(NSFC)

**References**


To Be or Not to Be: Will Scientific Writing Affect Scientific Impact?

Chao Lu¹ ²  Yi Bu²  Xianlei Dong³  Jie Wang¹  Ying Ding²  Chengzhi Zhang¹,*

luchaonjust@gmail.com;
¹ Nanjing University of Science and Technology, 200 Xiaolingwei Street, Nanjing, 210094 (China)

buyi@iu.edu;
Indiana University, 107 S Indiana Ave, Bloomington, IN, 47405 (U.S.A).

sddongxianlei@163.com;
Shandong Normal University, 88 Wenhua E Rd, Lixia Qu, Jinan, 250000 (China)

1342234559@qq.com;
Nanjing University of Science and Technology, 200 Xiaolingwei Street, Nanjing, 210094 (China)

dingying@indiana.edu;
Indiana University, 107 S Indiana Ave, Bloomington, IN, 47405 (U.S.A).

zhangcz@njust.edu.cn
Nanjing University of Science and Technology, 200 Xiaolingwei Street, Nanjing, 210094 (China)

Introduction
When it comes to possible benefit of linguistic complexity to scientific impact, opinions are split in communities (e.g., Gopen & Swan, 1990; O’Conner, 2010). To be or not to be complex? This is a worthy question for us to investigate—whether linguistic complexity is crucial to scientific writing so that impact can be expanded. Can scholars succeed when they write pleasant English, usually accompanied with complexity to some extent? Or they just employ writing as a tool to report their findings and ignore the beauty of language. After all, publishing papers of high impact as many as possible is one of scholars’ genuine concerns. This preliminary study is to dig this controversial questions based on large-scale fulltext and citation data of academic articles using regression analysis.

Methodology
Data
We collected 170,000 fulltext journal articles with publishing history (the dates when the paper is received, revised, accepted, and published) detailed in dates from 2006-2015 published in PLoS* and their corresponding citation data harvested from Scopus between 2016 February 3-6, a very short time period, so that we can neglect the potential error in citation data caused by different harvest timelines. We only kept all the articles in Biology by retrieving the assigned disciplinary information to reduce disciplinary differences. Our final dataset contained 49,350 fulltext articles in Biology with their publishing history and citation data with date.

Independent variables
The independent variable, Linguistic Complexity comprises Syntactic Complexity, the sentence-level complexity of language performance, and Lexical Complexity, the vocabulary-level language performance. Variables for syntactic complexity can be divided into two sub-groups: Sentence Length and Sentence Complexity (Vajjala & Meurers, 2012). And Lexical Complexity includes three sub-groups: Lexical Diversity, Lexical Sophistication, and Lexical Density (Ellis & Yuan, 2004; Kormos, 2011) (Table 1). Pearson’s Correlation Analysis showed no correlation between these variables.

Table 1. variables of linguistic complexity.

<table>
<thead>
<tr>
<th>Aspects</th>
<th>Indicators</th>
<th>Descriptions</th>
<th>Formulas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic Complexity</td>
<td>Sentence Length</td>
<td>Calculating average number of words in sentences and corresponding standard deviation of each article</td>
<td>$\frac{\sum_i N_i}{\sum_i T_i}$, $\frac{\sum_i N_i}{\sum_i V_i}$</td>
</tr>
<tr>
<td></td>
<td>Sentence Complexity</td>
<td>Counting the ratio of complex sentences that contain “that” or “which” in each article</td>
<td>$CR = \frac{\sum_i N_i}{\sum_i T_i}$</td>
</tr>
<tr>
<td></td>
<td>Lexical Density</td>
<td>Type-Token Ratio per 1000 words in each article</td>
<td>$\frac{TTR}{\text{words in each article}}$ × 1000</td>
</tr>
<tr>
<td></td>
<td>Lexical Diversity</td>
<td>Counting the ratio of lexical items in tokens in each paper based on their part of speech</td>
<td>$\frac{\sum_i M_i}{\sum_i T_i}$</td>
</tr>
<tr>
<td></td>
<td>Lexical Complexity</td>
<td>Counting the length of nouns, Sophistic verbs, adjectives, and adverbs on (4)</td>
<td>$\frac{\sum_i M_i}{\sum_i T_i}$, $\frac{\sum_i M_i}{\sum_i N_i}$, $\frac{\sum_i M_i}{\sum_i V_i}$</td>
</tr>
</tbody>
</table>

Dependent Variable
Since we had only obtained the total citation number for each article, we used the number of citations per month (CPM) as an alternative to eliminate the
possible effect caused by different periods of citation history. The variable was calculated as follows:

\[
CPM = \frac{\text{total number of citations}}{\text{Month(harvest date – published date)}}
\]

**Regression Analysis**

To investigate the relationship between Linguistic Complexity in scientific writing with scientific impact, we conducted regression analysis, using CPM as dependent variable and the 12 explanatory variables shown in Table 1. Considering that the variables had shown a strong trend of decrease after increase (two samples in Figure 1). Multinomial model \(Y = aX^2 + bX + c\) was to fit the data. articles with top 1% CPM ranking were selected as the most highly cited articles to compare the regression models.

![Figure 1. Scatter plots between CPM and JR (A) and MNL (B).](image)

**Results**

Table 2 shows descriptions of the two models: one for all the full text articles and another for the top 1% articles based on CPM ranking.

**Table 2. Descriptions of two regression models.**

<table>
<thead>
<tr>
<th>Variables</th>
<th>Model_All</th>
<th>Model_Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNL</td>
<td>1.504*** (0.111)</td>
<td>0.572*** (0.051)</td>
</tr>
<tr>
<td>MNL²</td>
<td>-0.0364*** (0.008)</td>
<td>-0.345*** (0.001)</td>
</tr>
<tr>
<td>MVL</td>
<td>0.698*** (0.152)</td>
<td>0.762*** (0.050)</td>
</tr>
<tr>
<td>MVL²</td>
<td>-0.0487*** (0.011)</td>
<td>-0.405*** (0.002)</td>
</tr>
<tr>
<td>MVL³</td>
<td>0.219** (0.069)</td>
<td>0.177** (0.047)</td>
</tr>
<tr>
<td>MVL⁴</td>
<td>-0.015*** (0.065)</td>
<td>-0.024*** (0.011)</td>
</tr>
<tr>
<td>TTR</td>
<td>-0.00265*** (0)</td>
<td>-0.0045*** (0.001)</td>
</tr>
<tr>
<td>TTR²</td>
<td>0.0000397*** (0)</td>
<td>0.00015*** (0)</td>
</tr>
<tr>
<td>CR</td>
<td>0.330*** (0.075)</td>
<td>0.302*** (0.018)</td>
</tr>
<tr>
<td>CR²</td>
<td>-0.487*** (0.118)</td>
<td>-0.502*** (0.032)</td>
</tr>
<tr>
<td>MSL</td>
<td>0.00775*** (0.002)</td>
<td>0.00855*** (0.011)</td>
</tr>
<tr>
<td>MSL²</td>
<td>-0.0000788* (0)</td>
<td>0.000135*** (0)</td>
</tr>
<tr>
<td>SSTD</td>
<td>0.00314*** (0.001)</td>
<td>0.00314*** (0.001)</td>
</tr>
<tr>
<td>SSTD²</td>
<td>-0.0000086*** (0)</td>
<td>-0.00058*** (0.001)</td>
</tr>
<tr>
<td>NR</td>
<td>0.970*** (0.085)</td>
<td>0.968*** (0.074)</td>
</tr>
<tr>
<td>NR²</td>
<td>-1.195** (0.5016)</td>
<td>-1.195** (0.5016)</td>
</tr>
<tr>
<td>VR</td>
<td>11.5*** (2.161)</td>
<td>11.5*** (2.161)</td>
</tr>
<tr>
<td>VR²</td>
<td>-36.24*** (7.482)</td>
<td>-36.24*** (7.482)</td>
</tr>
<tr>
<td>JR</td>
<td>6.184*** (0.86)</td>
<td>6.184*** (0.86)</td>
</tr>
<tr>
<td>JR²</td>
<td>-31.82** (5.385)</td>
<td>-31.82** (5.385)</td>
</tr>
<tr>
<td>VR²</td>
<td>53.91*** (5.385)</td>
<td>53.91*** (5.385)</td>
</tr>
<tr>
<td>cons</td>
<td>-6.154** (0.659)</td>
<td>25.72*** (8.994)</td>
</tr>
<tr>
<td>(R^2)</td>
<td>0.302</td>
<td>0.344</td>
</tr>
</tbody>
</table>

**Notes:** (1) Standard errors in parentheses; (2) \(p < .05, \quad **p < .01, \quad ***p < .001\).

In the first model, most features show quadratic relationships with CPM (p≤0.01), except for adverb length and noun ratio (positive linear relationship). The parameters suggest that moderate level of complexity in paper can promote the CPM and that the abundance of vocabulary provides a positive support for CPM of an article. In other words, moderate linguistic complexity is helpful to improving average citation of an article; however, too much complexity in syntactic level or word length may affect gain of citation since the difficulty in reading might dramatically increase according to studies (e.g., Juhasz, 2008).

However, the low level of R’s in the preliminary models might suggest that major efforts should be made in other areas to improve the content of an article, e.g., the novelty or contribution of the study articulated, which is supported by top CPM articles.

**Conclusion**

This preliminary study is to find out the relationship between linguistic complexity in scientific writing and scientific impact using regression analysis. The results might suggest that complex scientific writing helps improve scientific impact of an article to some extent but the major effort should be made in the content of it, e.g., the novelty and contribution to academic community. Given that the mutual effects of these variables to the impact has not been considered, more types of models should be used to clarify the relationships between them in the future.

**Acknowledgments**

This work is supported in part by Major Projects of National Social Science Fund of China (No. 16ZAD224).

**References**


1http://www.plos.org/
2Corresponding author: Chenghzi Zhang
A new method to quantify the spatiotemporal dynamic of the academic papers using centroid method

Wang Xuemei\(^1\)   Ma Mingguo\(^2\)

\(^1\)w20141103@swu.edu.cn
Southwest University Library, 2 Tiansheng Road, Chongqing (China)

\(^2\)mmg@swu.edu.cn
Southwest University, School of Geographical Sciences, 2 Tiansheng Road, Chongqing (China)

Introduction
The bibliometrics is an interdisciplinary field that combines mathematics, statistics and information science to study literature mining (Pritchard, 1969). The application of the geographic information system (GIS) technologies in the bibliometrics is an interdisciplinary field in the recent ten years (Allen, 2001). The spatial analysis capability is one of the most important functions in GIS. The application of GIS spatial analysis to bibliometrics is still in its initial stages (Wang et al., 2014). In this study, we try to expand the spatial analysis of the GIS in bibliometrics and develop a new method to quantify the spatiotemporal dynamic of the academic papers using centroid method.

Method
The global international papers of the geosciences from 1981 to 2016 were indexed and downloaded from the SCI-E and SSCI databases of the Core WOS Collection. Here the subject of the geosciences is based on the subject classification of Essential Science Indicators (ESI). The paper categories of ‘Article’ and ‘Review’ were included in our study. The published paper numbers and cited numbers of all authors for each country were counted.

For the polygon features, the geometric center is defined to balance the shape. The geometric center for each country was calculated based on this definition. The values of geographic parameter can be linked with the geometric centers of these polygon features. Then the centroid can be calculated based on the geometric centers and geographic parameters, which can reflect the change process of the spatial aggregation, dispersion and migration over time (Bansal et al., 2013). In this study, the published paper numbers and cited numbers of all the countries in the world were linked with these geometric centers. Then the centroid can be calculated using the following formulas.

\[
X_t = \frac{\sum_{i=1}^{n} (C_i \times X_i)}{\sum_{i=1}^{n} C_i} \quad (1)
\]

\[
Y_t = \frac{\sum_{i=1}^{n} (C_i \times Y_i)}{\sum_{i=1}^{n} C_i} \quad (2)
\]

Here, \(X_t\) and \(Y_t\) are the centroid coordinates of the longitude and latitude in the \(t^{th}\) year respectively. \(C_i\) is the published paper number or cited number of the \(i^{th}\) country in \(t^{th}\) year. \(X_i\) and \(Y_i\) are the longitude and latitude coordinates of the geometric center of the \(i^{th}\) country respectively.

Results
There are more than 190 countries or regions with the international papers of the geosciences during 1981 to 2016. But most of the countries or regions have the paper number less than 1000. Figure 1 shows that the percentage of the paper number for each country. The countries or regions with high percentages mainly distribute in the North America, Europe, Asia and Oceania. The top ten are United States of America (USA), United Kingdom (UK), China Mainland, Germany, France, Canada, Russia, Japan, Australia, and Italy in turn. The global international papers of the geosciences have the exponential increasing tendency and the exponential increasing coefficient is 0.0595 (\(R=0.9963\)) from 1981 to 2016. Similarly, most of the countries or regions have the exponential increasing tendency. The exponential increasing coefficients for the top 30 countries or regions with more paper numbers are shown in the Figure 1. It can be indicated that some countries or regions have higher coefficients than that of the global average level (green patches in Figure 1), such as South Korea, China Mainland, Chinese Taiwan, Turkey, Spain, Argentina, Austria, which means that they have more and more contributions on the global international papers of the geosciences from 1981 to 2016. Moreover, Russia, USA, Canada, New Zealand, UK, France, South Africa, Australia, India have the lower coefficients than that of the global average level (red patches in Figure 1).

The spatiotemporal dynamics of the centroids of the paper numbers and cited times are shown in Figure 2. It can be indicated that the centroids mainly move from west to east as a whole. The transfer speed of the centroids of the cited times is 1.43 degree/yr. Correspondingly, the transfer speed of the paper numbers is 0.93 degree/yr. It means that the cited times in the East have higher increasing speed than the paper numbers. On the
contrary, the centroids of the cited times are always located in the west of the centroids of the paper numbers. It indicates that the cited times in the West always occupy higher percentages than the paper numbers. But there is the decreasing trend from 1981 to 2016. The centroids of the paper numbers stuck at 17.7 degrees west of Greenwich from 1981 to 1995. On the contrary, the centroids of the cited times are approximately continuously migrating eastward.

![Figure 1. Exponential increasing coefficient and percentage of the paper number for each country](image1.png)

![Figure 2. The spatiotemporal dynamics of the centroids of the paper numbers and cited times.](image2.png)

**Conclusion**

The centroid method was successfully used to monitor the spatiotemporal dynamics of the paper numbers and cited times. Both of the centroids of paper numbers and cited times mainly move from west to east as a whole.

**Acknowledgments**

The study was supported by the Natural Science Foundation of China (Grant Nos. 41641058), Fundamental Research Funds for the Central Universities (XDJK2015C146).

**References**


Faculties Activity Research based on Local and Global Databases. Case Study.

Veslava Osinska¹ Piotr Malak²

¹ wieo@umk.pl
Nicolaus Copernicus University, Torun (Poland)
² poitr.malak@uwr.edu.pl
Wroclaw University, Wroclaw (Poland)

This research is sponsored by Polish National Science Center (NCN) under grant 2013/11/B/HS2/03048/Information Visualisation methods in digital knowledge structure and dynamics study.

Introduction

The state of under-representation of the humanities and social sciences in the global databases such as Web of Science and Scopus has become permanent and detrimental. But for scientometric research on mezzo (institutional) level local databases can successfully complement world indexes. The authors present case study of using unstructured data, deriving from University Website and then visually analysing. Cleaning and processing allows to identify metadata which describe academic activities in organizing different events at University. Another data source is University bibliographic database Expertus, the most complementary information resource concerning the publications of employees. This way, two different activities: publishing and organisational one can be compared for each faculty or equivalent division and thus evaluate collective impact in academic environment.

Data and metadata

Data collecting

Application form accessible at University Website https://www.umk.pl/badania/konferencje/ allows to collect such metadata about conferences as: Date, Title, Faculty, Organiser(s), Place, Leader, Secretary, Description. Downloaded data due to HTML format required semiautomatic cleaning. The relatively small records quantity $N = 1344$ allowed for quick estimation of similarities and relationships. Due to field about organizers units asymmetric matrix has been created for the quantitative representation of the intensity of cooperation in conference organizing. Asymmetry is connected with unequal roles of the organizer and co-organizer.

Data aggregation

All identified University units (over 300) were grouped into the main categories corresponding to 17 Faculties and one additional, General. Thus, for publishing and organisational activity, cumulative data according 18 categories was obtained. In order to compare the received visualisations with the generated image according to the Web of Science data, the next level of grouping: faculties into scientific areas was performed. Ultimately, into study we involved domains, such as:

- humanities,
- natural science,
- life sciences,
- medicine/clinical sciences,
- engineering and technical science,
- social sciences.

Analysis

Graphs, maps and diagrams were used in the visualisations. To illustrate the dynamics of publishing activity for each Faculty area chart is used as the most effective in the trend study (Fig. 1).

Figure 1. Dynamics of faculties publishing per year.

The band width exposes publication quantity a year for each faculty. It allows to observe all changes and parallel to capture information which is the most effective unit for analyzed period.

For comparison local and global databases, ring charts were used. Figure 2 presents domain structure for three data sources. As mentioned before, standardized categorization was applied. All domains colored by scheme, which is available at InCites reports: https://incites.thomsonreuters.com/. Outer ring at Fig. 2 shows the best representation of publications for humanity and social sciences according local database –Expertus.
Figure 2. Domain structure of publications of University employees according three databases.

The visualisation of social activities working on joint ventures (conferences, symposia, seminars) can be accomplished by using graphs (Garfield, 1994). We used the legible kind of graph – circular network layout (Fig. 3). Band width indicates the intensity of cooperation in organizing conferences. This way we discovered the close relations in such kind of activity between faculties (for example history and politology).

In order to prove this observation NLP techniques were used. Texts of conferences description were processed by Ward’s clusters analysis (1963) and visualised - the result we can see on Figure 4. Dendrogram presents faculties similarity based on texts clustering. Visualisation results confirmed previous assumptions according relationships between organizational units.

The set of metadata predisposes for extending mapping and broadening the perspective of study.

Figure 3. Co-organizing local events by different faculties.

We can also carry out geomapping and provide analysis of foreign cooperation of University. We can compare publishing versus mobility for each year, but all maps exceed the scope of current work.

Figure 4. Co-organizing local events by different faculties.

Summary

The authors present analytical possibilities of local database in confrontation with global scientometric indexes. Humanity and social sciences became visible on science map based on local data. Visualisations techniques were chosen according datatypes and analysis purposes (Osinska & Malak, 2016). Thus, circular graph shows cooperation between units, area chart – dynamics, ring diagram – comparison and structure and dendrogram – close and far similarity. Authors proposed publishing activity of scientists to complement by alternative one - organisational and this way to broaden evaluation framework.

References

Garfield, E. (1994) Scientography: Mapping the tracks of science. Social & Behavioural Sciences 7 (45), 5-10;
Always Gold Glitters:
When Accepted Papers Meet Processing Delay

Yingyi Zhang1 Chao Lu1,2 Chengzhi Zhang1,3,*
1. Department of Information Management, Nanjing University of Science and Technology, Nanjing, China, 210094
2. School of Informatics and Computing, Indiana University, Bloomington, US, 47405
3. Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University), Fuzhou, China, 350108
zyyzjs@163.com; luchaonjust@gmail.com; zhangcz@njust.edu.cn

Introduction
Processing delay is a time span between acceptance and publication (Shi et al., 2016). If an article is accepted, it is recognized by editor and reviewer (Shen et al., 2015), and processing delay is decided by the articles’ publish time. A long processing delay means articles are deposited in the editorial offices, when the value of articles can fade. Therefore, processing delay may pose a potential effect on the citation count of the article (Yu & Li, 2006). Given that citation count has become so dominant a metric to evaluate scholars’ research impact (Abbott 2009; King 2004), it is necessary to discover the potential relationship between processing delay and citation count. In order to clarify this question, this poster employs a correlation analysis between the two variables based on a large set of scientific articles published in PLoS, an influential open-access journal family. The results show that processing delay does not necessarily mean high citation count, and all gold will glitter regardless of processing delay.

Methodology

Data
We collected 170,000 full-text journal articles with publishing history (the dates when the paper is received, revised, accepted, and published) detailed in dates from 2006-2015 published in PLoS®. To reduce the influence of disciplines, we select articles whose disciplines is biology. The initial dataset contained 61,534 full-text articles in biology with their publishing history and citation data.

To strengthen the reliability of results, data was divided into subsets by the publish year (Letchford, 2015). Figure 1 shows the numbers of articles in different publication years.

\[ \text{Figure 1. Numbers of articles in different years} \]

Independent variables
Processing delay is calculated as follows by days:

\[ \text{ProcessingDelay} = \text{PublishDate} - \text{AcceptDate} \]  \hspace{1cm} (1)

Where, \( \text{PublishDate} \) represents when one paper is publish by a journal and \( \text{AcceptDate} \) represents when the journal accept it. For example, if \( \text{PublishDate} \) and \( \text{AcceptDate} \) of a paper is ’2008/11/14’ and ’2008/8/4’ respectively, then its processing delay will be 102 days.

Dependent Variable
Citation count measures how many times a paper is cited by others publications. To normalize citation count, this poster adopted the normalized citation count (NCC) formula developed by Uddin et al. (2012), which is a common practice. The formula is shown as follows stated:

\[ \text{NCC} = \frac{\text{CC}}{\text{Duration}} \times 365 \]  \hspace{1cm} (2)

Where \( \text{CC} \) stands for citation count. \( \text{Duration} \) indicates the number of days between the harvest date and the published date of the article. These citation data are harvested between 2016 February 3-6. Since the duration of data collection is short, so that we can mitigate the potential error in citation data caused by different harvest timelines.

* Corresponding author: Chengzhi Zhang.
1 http://www.plos.org/
Correlation analysis

To investigate the relationship between processing delay and normalized citation count, we conducted correlation analysis. The Kolmogorov-Smirnov (KS) test (p<0.001) shows that all variables are not normally distributed, so we cautiously adopted the Spearman test to discover whether there is some correlation between processing delay and normalized citation count.

Results Analysis

Correlation analysis results

Table 1 presents the correlation analysis results between processing delay and citation count from 2006 to 2014. From 2006 to 2012, the correlation between these two variables is not significant. In 2013 and 2014, the correlation is significant. However, these are all less than 0.1, which are very small correlations. Furthermore, the correlation of all is not significant. Since most of correlation tests are not significant and are very small correlations, processing delay and citation count do not have correlation. And fast processing delay does not necessarily mean high citation count.

<table>
<thead>
<tr>
<th>Year</th>
<th>correlation coefficients</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>0.061</td>
<td>0.594</td>
</tr>
<tr>
<td>2007</td>
<td>-0.020</td>
<td>0.599</td>
</tr>
<tr>
<td>2008</td>
<td>0.034</td>
<td>0.185</td>
</tr>
<tr>
<td>2009</td>
<td>-0.015</td>
<td>0.451</td>
</tr>
<tr>
<td>2010</td>
<td>-0.029</td>
<td>0.074</td>
</tr>
<tr>
<td>2011</td>
<td>0.011</td>
<td>0.299</td>
</tr>
<tr>
<td>2012</td>
<td>0.009</td>
<td>0.258</td>
</tr>
<tr>
<td>2013</td>
<td>-0.023</td>
<td>0.11</td>
</tr>
<tr>
<td>2014</td>
<td>0.030</td>
<td>0.000</td>
</tr>
<tr>
<td>All</td>
<td>-0.007</td>
<td>0.051</td>
</tr>
</tbody>
</table>

Note: All represents all the articles from 2006 to 2014.

![Figure 2](image)

Figure 2. Correlations of processing delay and citation count: (A) 2013 and (B) 2014.

Figure 2 represents the correlation of processing delay and citation count. Figure 2 (A) and (B) are samples of 2013 and 2014 respectively. Although there are strong trends of decrease in these figures, 2013 has negative correlation while 2014 has positive correlation in correlation tests. Trends of decrease are caused by some reasons including the long tail effect (Anderson, 2006) that most articles’ processing delay is short, and highly cited articles have high-probability of having a short processing delay.

Conclusion

This preliminary study is to analyse the correlation between processing delay and citation count. Results show that although processing delay can cause information loss, it is not the main factor of citation count. In order to strengthen the reliability of results, we also test the relationship in Medical Science; and we get similar results. Thus, scholars should not focus on how to shorten processing delays but the quality of their research (i.e., the novelty and contribution).

This poster did not control other variables (e.g. articles’ title and authors), and these variables should be taken in to consider in our following work. Furthermore, more articles of various discipline should be analysed.

Acknowledgments

This work is supported by Major Projects of National Social Science Fund (No. 16ZAD224), Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (No. MJUKF201704) and Qing Lan Project.

References


MOOC Content Design: Take “Webometrics & Scientific Evaluation” Course as an Example

Xiao Ming1 and Xu Ye2

1 ming_xiao@bnu.edu.cn
Beijing Normal University, Beijing(China)

2 xuye1219078794@qq.com
Beijing Normal University, Beijing(China)

Introduction
Massive Open Online Courses developed rapidly in 2012 and soon swept the world. MOOC has been accepted and promoted by the learners on the Internet because of its ease of use, low cost, rich learning resources, simplified video without time and space restrictions. How to design a popular content is a critical factor in determining the impact of the MOOC course. In this paper, taking “Webometrics & Scientific Evaluation” MOOC (as shown in Figure 1) as an example, we talk about our experience and understanding in designing the content of “Webometrics & Scientific Evaluation” MOOC, and hope it can be helpful when someone will design a MOOC in the future.

Methods
This paper uses the top-down design method to design the MOOC content. The whole MOOC content is divided into five parts: course objective, teaching goal, design principle, concrete plan and implementing scheme. Taking the course objective as the guidance, through the teaching goal and design principle of the norm, it will achieve concrete plan and implementing scheme, then complete the entire MOOC content design, as shown in Figure 2.

Course Objective
Course objective refers to the effect or impact that the MOOC is intended to achieve, such as curriculum style, social effect, teaching objects, etc.. The course objective of “Webometrics & Scientific Evaluation” MOOC is to create an interesting MOOC for the learners who are interested in Internet information resources, and hoping that people can improve their information literacy.

Teaching Goal
Teaching goal of the MOOC refers to the knowledge points that students should master, including the overall teaching goal of the MOOC course and the design of each section’s knowledge points and it’s difficulties. The teaching goal of “Webometrics & Scientific Evaluation” MOOC is to master the principle, method and application of webometrics. First, the knowledge points will be split into the introduction of webometrics, data sources of webometrics, analysis softwares of webometrics, analysis methods of webometrics, application of webometrics to scientific evaluation, and then refine these knowledge points step by step.

Design Principle
Design principle is the guideline of MOOC design, mainly including five principle: humanism-oriented education, practical content, streamline, fragmented but continuous, novel and diverse forms, as shown in Figure 3.

Concrete Plan
Concrete plan refers to the specific arrangements for the course content. It mainly solves following several questions. How many chapters will the MOOC course be divided into? How many sections will each chapter be divided into? How many knowledge points are there in each section? What specific material will be used for these knowledge points? Taking the Chapter 3, Section 3 of “Webometrics & Scientific Evaluation” MOOC as
Table 1 The concrete plan of “Webometrics & Scientific Evaluation” MOOC (Chapter 3, Section 3).

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Knowledge Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 3. Internet information resources can speak</td>
<td>Section 3. Web link analysis of web link relationship</td>
<td>3.3.1 “The life chain” of Internet</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.3.2 Web link analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.3.3 The index of web link analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.3.4 Basic steps for web link analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.3.5 Case analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3-2: Web link analysis of MOOC platform</td>
</tr>
</tbody>
</table>

Implementing Scheme
Implementing scheme refers to the specific design scheme about how to record MOOC videos according to the concrete plan. Formulate a clear scheme to display each knowledge point concretely, such as by animation, location shooting, interview, interaction with students, as shown in Figure 4 and Figure 5.

Results
By using a top-down approach, we first make a clear goal and plan well-ahead of time, and then refine and fill the MOOC content constantly. It can design a wonderful MOOC systematically and reach the initial course objectives effectively. This MOOC developed by Beijing Normal University (China) releases on 8 April 2017. At this time, we had more than 9000 students enrolled, as shown in Figure 6.

Acknowledgments
This work was supported by a grant from the National Social Science Foundation of China (No. 16BTQ073) and the online course project of Beijing Normal University (No. 02200-3122121J1).

References

Figure 6. The number of students enrolled and followers.
This MOOC provides examples and case studies for beginners that will stimulate creative solutions to the development and utilization of information resources on Internet. Moreover, this MOOC will benefit policymakers, practitioners, and managers who wish to connect with other professionals in order to enhance their knowledge of webometrics & scientific evaluation issues.

Figure 7. The reference materials of “Webometrics & Scientific Evaluation” MOOC.
Introduction

This paper describes the forecasting methods used to analyze the future demand for domain names under the ccTLD (country code Top-Level Domain) " .it". The aim of this analysis was to examine the trend over time of the demand for .it domain names. This is firstly to assess whether we are in a phase of growth and expansion or at a point of saturation. Secondly, this analysis enables us to compare the .it domain name market with those of other ccTLDs (.de, .uk, etc.) or of other gTLDs - general Top-Level Domains (.com, .edu, etc.).

Methods

In order to forecast the trend of demand for .it domain names, the main forecasting methods available in the literature were used. Before applying these forecasting methods, there was an analysis of the trend and seasonality of the series “variation in growth of .it” domain names. This is none other than the difference between registrations and cancellations recorded in the quarter x of year t. In general, a historical series can be broken down into a seasonal component, a trend and an erratic component. These components are usually estimated using the moving average. However, this method, albeit commonly used in the literature because of its ease of calculation, has some disadvantages. These are associated both with the limited accuracy of the estimates obtained and the loss of data relative to some terms due to the smoothing of the moving average. On the basis of these considerations, the trend, the seasonality and the erratic component of “ .it” domains were estimated using the loess method which, differently from the abovementioned method, gives more accurate estimates (Ricci, 2005). Subsequently, in order to estimate the forecasts of the .it domain names registration, some forecasting methods were applied. These included exponential smoothing methods, such as the Holt-Winters (additive and multiplicative) (Holt, 1957; Winters, 1960), which consider the presence of the trend and seasonality, and the stochastic processes like ARIMA - Auto Regressive Integrated Moving Average models (Box and Jenkins approach) (Box and Jenkins, 1976), in particular the SARIMA (seasonal ARIMA) version. The parameters of the various models were calculated using the maximum likelihood estimation. To assess the accuracy of the forecasts, resulting from the various forecasting methods, a range of indicators were used and analyzed which were able to give a measurement of the forecast error. These included the ME (Mean Error), MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), MPE (Mean Percentage Error), MAPE (Mean Absolute Percentage Error) and MASE (Mean Absolute Scaled Error). The latter indicator, was proposed by Hydman and Koehler (2006) as an alternative to using percentage errors when comparing forecast accuracy across series on different scales. Furthermore, due to the predictive accuracy of the methods considered, also other indicators were analyzed such as the Akaike Information Criterion (AIC), Corrected Akaike Information Criterion (AICc) and Schwarz Bayesian Information Criterion (BIC). Finally, in order to identify the best forecasting method, also the correlogram of the residuals of the various models were analyzed (ACF-Function of autocorrelation and PACF-Function of partial autocorrelation). The application of these forecasting methods and statistical indicators was carried out using the “R” statistical environment. The registrations and cancellations of .it domains were extracted in the .it domain names database managed by the Institute for Informatics and Telematics (IIT) of the CNR (National Research Council) in Pisa. The analysis was done taking into account the time period of nine years, from 2008 to 2016. Each year was subdivided into four quarters.

Results

From the analysis of the calculation of the trend and seasonal components, our time series shows a trend (slow growth during 2008, steadying between 2008 and 2009, reaching its maximum in the first half year of 2011, decreasing after 2011 until 2014, renewed growth to subsequently steady again). Furthermore, there is a seasonal component, given that the registrations and cancellations of .it domains were extracted in the .it domain names database managed by the Institute for Informatics and Telematics (IIT) of the CNR (National Research Council) in Pisa. The analysis was done taking into account the time period of nine years, from 2008 to 2016. Each year was subdivided into four quarters.
seasonal component, so as to make the series stationary, due to the seasonality in the demand for domain names. The choice of this model, compared to other ARIMA models, was made based on an analysis of the residuals and their correlograms (ACF and PACF).

Besides the ARIMA model, the other models applied to our series were both the Holt–Winters model with damped trend (Gardner and McKenzie, 1985) and additive seasonal component, and the Holt-Winters model with multiplicative trend and seasonal component. Also in this case the choice of these models instead of others was determined by considering the residuals and their correlograms. Figure 1 shows application of the abovementioned models to the .it domain name time series (registrations - cancellations).

As shown in figure 1, the ARIMA model fits our data better than the Holt-Winters models, above all in the first years (in particular in 2008 the values calculated by the ARIMA model are equal to the observed values).

While table 1 shows the forecasts that are produced by the three models, in the individual quarters of 2017 and 2018.

<table>
<thead>
<tr>
<th>Period</th>
<th>ARIMA</th>
<th>H-W multiplicative</th>
<th>H-W Additive Damped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 2017</td>
<td>54,995</td>
<td>52,379</td>
<td>48,920</td>
</tr>
<tr>
<td>Q2 2017</td>
<td>16,732</td>
<td>17,031</td>
<td>26,313</td>
</tr>
<tr>
<td>Q3 2017</td>
<td>32,776</td>
<td>29,968</td>
<td>24,596</td>
</tr>
<tr>
<td>Q4 2017</td>
<td>30,470</td>
<td>32,776</td>
<td>24,596</td>
</tr>
<tr>
<td>Q1 2018</td>
<td>54,781</td>
<td>54,201</td>
<td>39,065</td>
</tr>
<tr>
<td>Q2 2018</td>
<td>16,613</td>
<td>17,624</td>
<td>25,941</td>
</tr>
<tr>
<td>Q3 2018</td>
<td>32,709</td>
<td>31,010</td>
<td>24,299</td>
</tr>
<tr>
<td>Q4 2018</td>
<td>30,433</td>
<td>38,523</td>
<td>38,827</td>
</tr>
</tbody>
</table>

Discussions and Conclusions

The estimates of the parameters of the Holt-Winters damped additive model indicate that the estimate of the level at the current time point is based on the less recent observations (the less recent observations have a greater weight than the more recent ones). On the other hand, for the Holt-Winters multiplicative model the estimate of the level at the current time point is based not only on the less recent observations but also on the most recent ones.

From analysis of the forecasts, the ARIMA model is the one which better fits the historical data compared with the models of exponential smoothing. This conclusion is valid, both when observing the graphs relative to the diagnostics of the residuals and the statistical indicators relative to the accuracy of the estimates. In fact, all the statistical indicators that give a measurement of the error of the estimates (ME, MAE, RMSE, MPE, MAPE and MASE) are lower when applying the ARIMA model, compared to the other two models.

As regards the diagnostics of the residuals, the p values for the Ljung-Box statistic in the various lags are higher for the ARIMA model, indicating that the hypothesis of absence of autocorrelation of the residuals, for all the lags, cannot be rejected. On the contrary, for the Ljung-Box statistic of the Holt-Winters models, the p values are very low compared to the ARIMA model and in some cases the test rejects the hypothesis of uncorrelated residuals. The same applies to other indicators which give a measure of the degree of fit of the models used. AIC, AICc and BIC are much lower for the ARIMA model than for the models of exponential smoothing.

In conclusion, according to the results in Table 1, the .it domain name registrations will be greater than the cancellations in both 2017 and 2018. For example, in 2018 it is foreseen that there will be an annual positive growth in domain names by about 135,000 domains as forecast by applying the ARIMA model, and by approximately 141,000 and 138,000 by applying the Holt-Winters multiplicative model and the Holt-Winters damped additive model respectively.

In summary, notwithstanding the forecast of a growth in domain names in the 2017-2018 two-year period, the increase is however limited, tending to reach a phase of saturation of the “.it” market.

References


Are the social sciences from the European post-socialist countries integrated in the "Western social sciences"?

Maja Jokić¹ Stjepan Mateljan¹ Nikola Petrović¹

maja@idi.hr; smateljan@idi.hr; nikola@idi.hr

Institute for Social Research in Zagreb, Amruševa 8, 10 000 Zagreb (Croatia)

Introduction

There are significantly fewer bibliometric studies focusing on social sciences than those in the field of natural sciences. Reasons for this vary from a conventional opinion that the paradigm of science communication in social sciences is significantly different from that in STEMM areas and the perception of social sciences as particularly susceptible to limitations of bibliographic and citation databases (Arunachalam et al., 2006; Nederhof, 2006; Butler & McAllister, 2009; Ossenblok & Engels, 2015) to a wide array of complex features within social sciences such as cultural and linguistic differences, national traditions, state-centred institutionalisation and locally-oriented research themes (Mosbah-Natanson & Gingras, 2014). According to Garneau (1985), until the end of the 1980s social sciences were divided into Western social sciences, or ‘first world social science’, and social sciences from communist countries, or ‘second world social science’. However, today’s European Research Area (ERA) includes all central and eastern European post-socialist countries (CEE), either as EU member states or EU candidate countries, and one of its principal aims is a complete integration mentioned countries into the European Union Research and Development Programs (Bruhns, 2012). In this research we tried to establish the extent to which social sciences in European post-socialist countries (N=15; 11 EU member countries and 4 EU candidate countries) have so far been integrated into ERA, using publishing and citation data for international/CEE journals.

Sample and Methodology

The sample was obtained from the Scopus database and it consists of papers in the field of social sciences authored by at least one researcher from one of 15 European post-socialist countries – 11 EU members states, i.e. Estonia, Latvia, Lithuania, Poland, Czech Republic, Slovakia, Hungary, Romania, Bulgaria, Slovenia, Croatia, and 4 EU candidate countries, i.e. Bosnia and Herzegovina, Macedonia, Montenegro and Serbia. The analysed papers (N=33,837), published in the period 1996 - 2013 in 2,587 international and 157 CEE journals, were reclassified according to the Croatian science classification, which is based on OECD classification fields (which include economics, education, information and communication science, law, political science, psychology, sociology and three interdisciplinary fields). We analysed paper distribution in international/CEE journals, paper frequency by subject fields and CEE countries' research output, while in case of international journals we analysed publishing orientation by country.

Results

Apart from the expected rise in CEE countries' research output, which is visible due to a relatively high number of CEE journals included in the Scopus database, the results indicate significant increase in the publication of papers in international journals after 2004 (Fig. 1), with their total share being 56%.

Figure 1. Paper distribution in international/CEE journals between 1996 and 2013 (Source: Scopus).

European post-socialist countries predominantly published their research results in social sciences in Western countries journals (Fig. 2). UK’s, Netherlands’ and Germany’s journals represent 90% of journals from EU-15 countries.
In both CEE and international journals, the majority of papers are in the field of economics, with their total share being 32.5%. However, scientific fields showing the strongest international orientation are political science, information science and all interdisciplinary social science fields (Fig 3).

The normalisation of each of the sampled countries’ research output by their respective population numbers shows that Estonia, Slovenia, Hungary and Lithuania are the most internationally oriented countries (Fig. 4).

One of the aspects of measuring integration is citedness which in international journals is 7.4 citations per paper, and in CEE journals 2. For a more comprehensive insight there is a need for more detailed citation analyses.

Conclusion

The results show evidences of a gradual integration of social sciences in European post-socialist countries into “Western social sciences”. A significant increase in this trend occurred after 2004, when most of CEE countries joined the EU. Political science, information science and interdisciplinary social science are fields that are frontrunners in this integration. CEE countries that are most oriented towards international journals are small CEE countries and Hungary. The project is planning more detailed bibliometric analysis, co-citation analysis as well as analyses of social contexts which all have a significant influence on integration.

Acknowledgments

This work has been supported by Croatian Science Foundation under the project IP-09-2014-9351.

References:
An Artificial Neural Network Model Based on Altmetrics Indicators and Citation Counts

Li Xiaotao\textsuperscript{1}  Chen Changhua\textsuperscript{2}  Qin Ping\textsuperscript{3}

\textsuperscript{1}lixiaotao@nuaa.edu.cn
\textsuperscript{2}cchlib@nuaa.edu.cn
\textsuperscript{3}767085738@qq.com

Library, Nanjing University of Aeronautics and Astronautics, NO. 29 Yudao Road, Qinhui District, Nanjing, Jiangsu Province (China)

Introduction

The main advantages of altmetrics over traditional bibliometrics and webometrics is that they offer fast, real-time indications of impact, they are openly accessible and transparent, include a broader non-academic audience, and cover more diverse research outputs and sources (Costas et al. 2015). In the last few years, the interest in altmetrics has grown, giving rise to many questions regarding their potential applications.

The relationship between citation counts and altmetrics indicators of academic papers was a research hotspot in the field of informetrics and scientometrics. A meta-analysis across more than 40 cross-metric validation studies shows overall a weak correlation (ranging from 0.08 to 0.5) between altmetrics and citation counts, confirming that altmetrics do indeed measure a different kind of research impact, thus acting as a complement rather than a substitute to traditional metrics (Erdt et al. 2016). The quantitative relationship and influence mechanism between citation counts and altmetrics indicators required further research to reveal. Shall we predict papers citation counts with its altmetrics indicators?

Artificial neural networks were widely used for building the nonlinear model and pattern recognition among a number of variables(Basheer and Hajmeer, 2000). This paper aims to reveal the quantitative relationship and influence mechanism of citation counts and altmetrics indicators (such as bloggers, tweeters, F1000 reviews, Facebook walls, and so on) by constructing an artificial neural network model. The model’s veracity will be verified.

Material & Methods

PlumX, ImpactStory, and Altmetric.com were usually used for collecting data of altmetrics indicators(Peters et al. 2016). In this study, we select Web of Science and Altmetrics.com as data sources to collect citation counts an altmetrics indicators of academic papers. On March 15th of 2017, We respectively searched papers published on Nature in 2015 from the two data sources. Papers from the two sources were matched by their Digital Object Identifier (DOI). 968 papers were matched, then we got citation counts and altmetrics indicators of these papers.

The training set included 726 papers and their indicators, while the test set included the rest of the papers. We chose 11 indicators to construct the artificial neural network, including 10 altmetrics indicators (Bloggers, Tweeters, Google+ authors, F1000 reviews, News outlets, Facebook walls, Weibo users, Peer review sites, Wikipedia pages, Mendeley readers, CiteULike readers) as input node, and papers’ citation counts as output node.

The artificial neural network was constructed by R-project. All data had been normalized before imported into R. The correlation between predicted citation counts by the model and the actual value of test sets was analysed to evaluate the model precision.

Result & Discussion

Data collection

We collected citation counts and altmetrics indicators of 968 papers published on Nature from web of science and Altmetrics.com respectively. There were 11 indicators selected for constructing artificial neural networks, including bloggers, tweeters, google+ authors, F1000 reviews, news outlets, facebook walls, weibo users, wikipedia pages, mendeley readers, CiteULike readers and citation counts. Some of the data were listed in Table 1.

Data normalization

Due to the numerical distribution range of indicators was not good for the construct of artificial neural network, all data had been scaled by min-max normalization before imported into R-project.
\[ y_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \]

Table 1. Altmetrics indicators and citation counts of papers.

<table>
<thead>
<tr>
<th>Index</th>
<th>Paper1</th>
<th>Paper2</th>
<th>Paper 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bloggers</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Tweeters</td>
<td>42</td>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>Google+</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>F1000</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>News Outlets</td>
<td>1</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Facebook</td>
<td>2</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Weibo</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mendeley</td>
<td>332</td>
<td>161</td>
<td>116</td>
</tr>
<tr>
<td>CiteULike</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Citation</td>
<td>52</td>
<td>75</td>
<td>96</td>
</tr>
</tbody>
</table>

**Model construction**

The neuralnet package of R was used to construct the artificial neural network model. The training set included 726 papers. Altmetrics indicators were trait as input node, while citation counts were considered as output node and two hidden node were inserted. The model constructed on training set was shown as figure 1.

**Table 1. Altmetrics indicators and citation counts of papers.**

![Model construction](image)

**Figure 1. The artificial neural network model**

**Model precision**

We imported data of test set including 242 papers into the model and calculated citation counts by data of input nodes (altmetrics indicators). The correlation coefficient of predicted result and actual citation counts was 0.62. The Scatter diagram of predicted result and actual citation counts was shown in figure 2.

![Figure 2. Scatter diagram of predicted result and actual citation counts.](image)

**Conclusion**

The artificial neural network is highly nonlinear, capable of complex logical operation and non-linear relationship realization. The artificial neural network provides a new way to explore the quantitative relationship between altmetrics indicators and citation counts. The artificial neural network model of this paper is only a preliminary attempt, and the complexity and accuracy of this model is still to be optimized. In the future, we may be able to construct more accurate mathematical models with deep neural networks, and predict the cited frequencies of papers by altmetrics indicators.

**References**


Using Full-text to Evaluate Impact of Different Software Groups

Shutian Ma\textsuperscript{1}  Chengzhi Zhang\textsuperscript{1,2,*}

\textsuperscript{1}Department of Information Management, Nanjing University of Science and Technology, Nanjing, China, 210094
\textsuperscript{2}Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University), Fuzhou, China, 350108

mashutian0608@hotmail.com; zhangcz@njust.edu.cn

Introduction
To analyse the impact of software on science, researchers have tried to extract software entities from scientific literature (Pan et. al., 2015) and examined how software is mentioned and cited across disciplines (Pan et. al., 2016). However, previous study doesn’t consider functional relations among different software. In this paper, we evaluate impact of software groups with the same usage. Firstly, deep learning models are used to learn distributed representations of software entities based on full-text. Clustering technique is then applied to find software groups. Finally, we analyse the impact of each software group.

Methodology

Framework
Figure 1 shows framework of impact evaluation of different software groups.

\begin{figure}[h]
  \centering
  \includegraphics[width=\textwidth]{framework.png}
  \caption{Frameworks of Impact Evaluation of Different Software Groups}
  \label{fig:framework}
\end{figure}

Firstly, we represent each software as vectors via distributed representation. Then, we divide them into different groups by clustering method. Finally, we count the mentioned times of each software and evaluate software groups based on total number of mentions and average number of mentions in full-text.

Distributed representation
Currently, Word2Vec (Mikolov et. al., 2013) can provide state-of-the-art word embedding. It can learn high-quality word vectors from huge corpora automatically. Based on this, Doc2Vec (Para2Vec) is proposed to obtain representations for larger blocks of texts, such as sentences, paragraphs even entire documents (Le & Mikolov, 2014). Specially, we add procedure which matches software entities and groups literature into documents set containing the same software when training Doc2Vec model. Here, word dimension is 200, document dimension is 100. The maximum distance between predicted one and context used for prediction is 5.

Clustering algorithm
Affinity propagation (Frey & Dueck, 2007) is used for clustering. It works based on similarity between pairs of all the data points by considering all points as potential cluster centres. In this experiment, preference is needed to be set which indicates preference that data point is chosen as a center. Here, similarity between documents are calculated with squared Euclidean distance and preference is set up to be: -5, -10, -15, -95, -100, with 5 as interval.

Experiment and Results Analysis

Dataset and Tools
The scientific literature are all collected from PLOS ONE\textsuperscript{1}. We downloaded 114, 510 papers in XML format. Software list is obtained from work (Pan et. al., 2015). We clustered 1, 617 software entities. Word2Vec and Doc2Vec are applied in Genism\textsuperscript{2}. Affinity propagation is done via Scikit-learn\textsuperscript{3}.

Evaluation Metrics
Silhouette Coefficient (Rousseeuw, 1987) is used to measure clustering performance. Suppose a is the mean distance between a sample and all other points

\footnotesize{\textsuperscript{1}http://journals.plos.org/plosone/}
\footnotesize{\textsuperscript{2}http://radimrehurek.com/gensim/index.html}
\footnotesize{\textsuperscript{3}http://scikit-learn.org/stable/index.html}
in the same class, $b$ is the mean distance between a sample and all other points in the next nearest cluster. It’s calculated via formula below:

$$Sil = \frac{b-a}{\max(a, b)}$$

(1)

Experimental Results Analysis

Clustering results which cluster number is over than 4 and silhouette coefficient is above 0 are shown in table 1 and 2. Evaluations are made based on the two best results shown in bold italics. A higher silhouette coefficient means better clustering results.

Table 1. Clustering results of software represented by Word2Vec

<table>
<thead>
<tr>
<th>No.</th>
<th>All Men</th>
<th>Ave</th>
<th>Exemplar</th>
<th>Cluster members (Top5)</th>
<th>Clu</th>
<th>Sil</th>
<th>Pre</th>
<th>Clu</th>
<th>Sil</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>127,865</td>
<td>110</td>
<td>KING</td>
<td>SPSS, Prizm, serial, SAS, Survival</td>
<td>Pre</td>
<td>-10</td>
<td>0.93580</td>
<td>-10</td>
<td>0.29395</td>
</tr>
<tr>
<td>2</td>
<td>39,756</td>
<td>128</td>
<td>MAGIC</td>
<td>Cluster, Flojo, Vienna, PASW, MACS</td>
<td>6</td>
<td>0.40460</td>
<td>-5</td>
<td>45</td>
<td>0.17247</td>
</tr>
<tr>
<td>3</td>
<td>5554</td>
<td>1386</td>
<td>GTC</td>
<td>ACT, GGT, GTC, T-Rex</td>
<td>-85</td>
<td>5</td>
<td>0.04820</td>
<td>-25</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>7</td>
<td>3D-MSPECT</td>
<td>QGIS, 4D-MSPECT</td>
<td>-30</td>
<td>8</td>
<td>0.04823</td>
<td>-5</td>
<td>54</td>
</tr>
</tbody>
</table>

Table 2. Clustering results of software represented by Doc2Vec

<table>
<thead>
<tr>
<th>No.</th>
<th>All Men</th>
<th>Ave</th>
<th>Exemplar</th>
<th>Cluster members (Top5)</th>
<th>Clu</th>
<th>Sil</th>
<th>Pre</th>
<th>Clu</th>
<th>Sil</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>51,047</td>
<td>193</td>
<td>ModFit</td>
<td>BLAST, Cluster, Flojo, Stata, Eclipse</td>
<td>Pre</td>
<td>-20</td>
<td>0.93580</td>
<td>-10</td>
<td>0.29395</td>
</tr>
<tr>
<td>2</td>
<td>12,808</td>
<td>61</td>
<td>MasterPl ex</td>
<td>EXCEL, Access, SPM, TargetScan, Sequencher</td>
<td>6</td>
<td>0.40460</td>
<td>-5</td>
<td>45</td>
<td>0.17247</td>
</tr>
<tr>
<td>3</td>
<td>36,750</td>
<td>168</td>
<td>GENEMAP</td>
<td>SPSS, MEGA, BigDye, JMP, Pfam</td>
<td>-85</td>
<td>5</td>
<td>0.04820</td>
<td>-25</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>16,146</td>
<td>158</td>
<td>ANSLAB</td>
<td>ImageJ, PASW, StepOne, FSL, Scaffold</td>
<td>-30</td>
<td>8</td>
<td>0.04823</td>
<td>-5</td>
<td>54</td>
</tr>
<tr>
<td>5</td>
<td>14,461</td>
<td>96</td>
<td>ImPerm</td>
<td>Survival, Platinum, MrBayes, AxoVision, HiSeq</td>
<td>Pre</td>
<td>7</td>
<td>0.93580</td>
<td>-10</td>
<td>0.29395</td>
</tr>
<tr>
<td>6</td>
<td>32,140</td>
<td>150</td>
<td>adze</td>
<td>SDS, SAS, Vienna, SAM, LSR1</td>
<td>5</td>
<td>0.40460</td>
<td>-5</td>
<td>45</td>
<td>0.17247</td>
</tr>
<tr>
<td>7</td>
<td>30,963</td>
<td>184</td>
<td>SerialEM</td>
<td>Prizm, Adobe, Chromosome, Cycler, RMA</td>
<td>Pre</td>
<td>6</td>
<td>0.93580</td>
<td>-10</td>
<td>0.29395</td>
</tr>
<tr>
<td>8</td>
<td>46,975</td>
<td>163</td>
<td>NEWBLE</td>
<td>serial, Agilent, geometry, ACT, Ingenuity</td>
<td>Pre</td>
<td>5</td>
<td>0.93580</td>
<td>-10</td>
<td>0.29395</td>
</tr>
</tbody>
</table>

Figure 2(a) displays Word2Vec results where we can find two clusters directly, figure 2(b) displays results of Doc2Vec. As we can see, software are projected into different places using Word2Vec and Doc2Vec.

Software entities of each clusters are also shown in table 3 and 4. We give top-5 entities ordered by number of mentions. All Men is the all the software mentioned times within each cluster, Ave is the average mentioned times. Firstly, we find software within cluster have the same usage, like analysis of sequence data, graphics, visualization and so on. For example, in table 3, top-5 members of 6th cluster are software to analyse biological sequences, which is the same with 3rd cluster in table 4. Based on number of all mentions in table 3 and average mentions in table 4, statistical software for gene data tend to have more impact than others. Besides, it shows our method can help us to find similar software by clustering.

Figure 2. Visualization of Clustering Results

Conclusion

In this paper, we evaluate the impact of different software groups based on full-text. Experimental results show that statistical software shows more impact than other kind of software in PLOS ONE literature. Current conclusion is simply from the preliminary investigation, problems are needed to be solved, like cluster analysis, matching strategies.

Acknowledgments

This work is supported by Major Projects of National Social Science Fund (No. 16ZAD224), Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (No. MJUKF201704) and Qing Lan Project.

References


Specialization and excellence in Latin American and Caribbean science: assessment proposal

Dirce Maria Santin¹ and Sônia Elisa Caregnato²

¹ dirce.santin@ufrgs.br
Universidade Federal do Rio Grande do Sul, Rua Sarmento Leite, 500 – sala 205, CEP 90050-170, Porto Alegre (Brazil)

² sonia.caregnato@ufrgs.br
Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos, 2705 – sala 212, CEP 90035-007, Porto Alegre (Brazil)

Introduction
Monitoring research activity and impact is essential for the competitiveness of countries and regions since it can indicate strengths in science and identify fields that need to improve performance in the global context.

This paper presents an assessment proposal of regional and global science in Latin America & Caribbean (LAC) based on bibliometric indicators and contributions of Sociology of Science. The project aims to analyze the scientific profiles of countries in the region, their strength, excellence and scientific specialization, and the reflexes of these configurations on the positions occupied by the countries in global science.

The research focus is on five dimensions: (1) countries’ publication profile and impact; (2) countries’ scientific performance (scientific strength); (3) scientific specialization of the region and its countries (activity and impact normalized); (4) identification of centers of excellence in regional and global science (highly cited publications); and (5) reflexes of these configurations on the positions occupied by these countries in the international scientific field.

Data sources and methodology
The present study is based on bibliographic data extracted from Web of Science - WoS (Science Citation Index, Social Science Citation Index and Arts & Humanities Citation Index) and SciELO Citation Index. Only original and review articles published between 2003 and 2014 were taken into consideration. Publications were assigned to countries on the basis of their corporate addresses which appear in the field country (CU) of WoS. All countries of LAC listed in the Standard Country or Area Codes for Statistics Use of United Nations and indicated in the address field were considered. A three-year citation window was applied and included citations and self-citations. The full counting method was used and the thematic classification is based on research areas (SC), common in the two databases. In macro analyzes the SC were transposed into the six major fields of the Frascati Manual.

Data collection was carried out in March 2017 and the search strategy combined the fields country (CU), document type (DT), and publication year (PY). The data were collected in four steps:

Step 1. Bibliographic records of LAC articles and absolute citations.
Step 2. Global absolute numbers of articles by SC.
Step 3. Citation data of region, by year and SC.
Step 4. Citation data by country, year and SC.

The collected data was pooled into two unique databases (WoS and SciELO) and the duplicate records were excluded. The statistical language R (RStudio interface), Microsoft Excel and BibExcel were used for data processing and analysis.

Publication profiles will be assessed by means of Activity Index (AI) and Country Profile Index (CPI), while the citation profiles will be based on the Relative Citation Impact Score (RCIS) and Attractivity Index (AAI) (Schubert & Braun, 1986; Tuzi, 2005; Schulz & Manganote, 2012).

AI = \frac{\text{Percentage of country publications in the field}}{\text{Percentage of world publications in the field}}

CPI = \frac{\text{Number of country publications in the field}}{\text{Number of country publications in all fields}}

RCIS = \frac{\text{Average of country citations in the field}}{\text{Average of regional citations in the field}}

AAI = \frac{\text{Percentage of country citations in the field}}{\text{Percentage of country citations in all fields}}

The countries specialization profiles will also be evaluated with the indicators Scientific Strength (SS) (Abramo, D’Angelo, & Di Costa, 2015) and Scientific Specialization Index (SSI) (Abramo, D’Angelo, & Di Costa, 2014), which consider both publications and citations received. SS results from the sum of the normalized impact, obtained by Article Impact Index (AII), and is relativized for the total field values in the LAC.

AII = \frac{\text{Number of citations of article in field}}{\text{Average citations of articles cited in field}}
The SSI is based on the IIA and SS values, and integrates the data normalized by field, country and year of publication. The SSI of the country k in the SC j \((SSI_{kj})\) is defined as:

\[
SSI_{kj} = 100\cdot \tanh \left( \frac{SS_{kj}}{\sum_{k}SS_{kj}} \left( \frac{\sum_{k}SS_{kj}}{\sum_{k}SS_{kj}} \right) \right)
\]

Identification of centers of excellence will be based on highly cited articles by authors from the region, starting from a threshold of 5% of the most cited articles in the WoS and 1% in the SciELO. This avoids a limit of citations for each research area which could exclude most of the fields and compromise the analysis of the geographical distribution of centers of excellence (Tijssen, Visser, & van Leeuwen, 2002). Positions held by LAC countries in the international scientific field (Bourdieu, 1988) will be evaluated by aspects such as: positions of countries in world rankings (outputs and impact); participation and leadership of international organizations; awards; among others.

**Preliminary results**

The LAC output in WoS in the period is 643,221 articles, with a mean of 12.8 citations per article and a total of 8,231,334 citations. In the SciELO database the region presents 274,335 articles and 513,903 citations – mean of 1.9 citations per article. The study includes comparative analyzes between global or mainstream science (WoS) and regional or peripheral science (SciELO Citation Index). This paper, however, presents only preliminary results based on WoS data, without results of SS and SSI. Figure 1 shows the evolution of LAC output by major fields of the Frascati Manual.

**Figure 1. Evolution of LAC output by major field in WoS, 2003-2014**

The pattern of linear growth in the number of articles reaches a certain degree of stability in recent years. Natural Sciences and Medical & Health Sciences are the most productive fields of LAC, with \(R^2\) 0.9944 and 0.9618, respectively. Table 1 presents the relative indicators of activity and impact of the most productive LAC countries in the three SC with the highest number of articles.

**Table 1. Activity and impact of the most productive LAC countries in WoS, 2003-2014**

<table>
<thead>
<tr>
<th>Country</th>
<th>SC</th>
<th>AI</th>
<th>CPI</th>
<th>RCIS</th>
<th>AAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>Chemistry</td>
<td>0.17</td>
<td>0.09</td>
<td>1.02</td>
<td>1.01</td>
</tr>
<tr>
<td>Agriculture</td>
<td>4.51</td>
<td>0.09</td>
<td>0.83</td>
<td>1.10</td>
<td></td>
</tr>
<tr>
<td>Physics</td>
<td>0.21</td>
<td>0.09</td>
<td>1.06</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>Mexico</td>
<td>Chemistry</td>
<td>0.06</td>
<td>0.10</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>Engineering</td>
<td>0.07</td>
<td>0.09</td>
<td>0.91</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>Argentina</td>
<td>Chemistry</td>
<td>0.05</td>
<td>0.12</td>
<td>1.04</td>
<td>1.05</td>
</tr>
<tr>
<td>Physics</td>
<td>0.06</td>
<td>0.11</td>
<td>1.19</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>Biochemistry &amp; Mol. Biology</td>
<td>0.13</td>
<td>0.06</td>
<td>1.08&amp; 1.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chile</td>
<td>Astronomy &amp; Astrophysics</td>
<td>2.28</td>
<td>0.12</td>
<td>1.29</td>
<td>0.82</td>
</tr>
<tr>
<td>Physics</td>
<td>0.03</td>
<td>0.08</td>
<td>1.21</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>Chemistry</td>
<td>0.02</td>
<td>0.07</td>
<td>0.92</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>Colombia</td>
<td>Engineering</td>
<td>0.03</td>
<td>0.14</td>
<td>0.65</td>
<td>1.19</td>
</tr>
<tr>
<td>Chemistry</td>
<td>0.01</td>
<td>0.08</td>
<td>0.84</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>Physics</td>
<td>0.02</td>
<td>0.12</td>
<td>1.45</td>
<td>1.44</td>
<td></td>
</tr>
<tr>
<td>Venezuela</td>
<td>Chemistry</td>
<td>0.01</td>
<td>0.11</td>
<td>0.92</td>
<td>1.01</td>
</tr>
<tr>
<td>Physics</td>
<td>0.01</td>
<td>0.09</td>
<td>0.77</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>Engineering</td>
<td>0.01</td>
<td>0.08</td>
<td>0.86</td>
<td>0.98</td>
<td></td>
</tr>
</tbody>
</table>

The most productive areas are not necessarily those with greater activity and relative impact. This suggests that the absolute productivity, by itself, does not reveal whether the country is more or less active or specialized in the field, nor its contribution to global knowledge.

**Acknowledgments**

We are grateful to the ISSI Student Travel Award and Universidade Federal do Rio Grande do Sul.

**References**


On normalization by division and comparing citation counts

Lawrence Smolinsky¹

¹smolinsk@math.lsu.edu

Department of Mathematics, Louisiana State University, Baton Rouge, LA (USA)

Introduction

Bibliometric measures based on citation counts are often used as an objective measure of research productivity. Finding a method of comparison is extremely challenging given the diversity of scholarly publications particularly by field and age. These are some aspects that effect the way publications accumulate citations. Developing measures using citations depends on understanding the distributions of citations of publications. Methods for normalizing the count across fields for comparison is in constant development.

Division

One heuristically attractive idea is to divide by a version of the expected number of citations to the paper. It may be done for individual papers, institutions, states, etc. Some of the reasons division is attractive is that it is simple to do and transparent in that percentage above or below average may be used in daily life. Some indicators that rely on division are the mean normalized citation score (MNCS) or InCites’s mean category normalized citation impact or normalized citation impact.

Whether division by a mean, median, or other parameter can make citation counts between different areas comparable is dependent upon the distributions being normalized. Each field may be thought to have its own distribution of the number of papers having a given number of citations. These distributions can be interpreted as probability distributions by dividing by the number of papers. For the purpose of cross field comparisons, they should ideally be adjusted so they have the identical distributions for each field. Normalizing by dividing by the mean may give the various distributions the same mean, but otherwise they may be quite different. However, there are families of distributions which become identical with a division adjustment by the mean. In other words, there are well-known families of distributions for which two random variables X and Y in the family with respective distinct means mx and my have \( \frac{x}{mx} \) and \( \frac{y}{my} \) identically distributed. The mean is then a scale parameter. These distributions include the exponential, Exp(\( \lambda \)), the gamma distribution with fixed \( n \), \( \Gamma(n, \lambda) \), and, a standard candidate in bibliometrics, the lognormal with a fixed \( \sigma \), Lognormal(\( \mu, \sigma^2 \)). Discrete distributions are problematic for scaling since the domain may change.

The lognormal scaling was essentially examined in Radicchi, Fortunato, and Castellano (2008) for data in sample areas for the year 1999. They examined the distribution of articles published in a single year in a specific area and approximated it by a lognormal and then compared the various distributions by scaling. The model has its flaws. For example, the lognormal is zero at zero while the number of articles with zero citations is significant. Furthermore, while Radicchi, Fortunato, and Castellano only give an exponential scale graph including multiple areas (2008 Fig. 2 p. 17269), one can visually see some fields below the normalized graph in the tail, e.g., anesthesiology and mathematics. However, suppose one accepts the lognormal with a fixed parameter \( \sigma \) for as the model for the distribution of citations in all fields, i.e., only the parameter \( \mu \) changes from field to field. Then division by the mean for each field normalizes the citation distribution to the common distribution Lognormal[-\( \sigma^2/2 \), \( \sigma \)] for each field. An indicator like the MNCS, which is a sum of \( n \) normalized random variables (for a fixed \( n \)), will have a common distribution across fields. Even though the distribution may be complicated, it will the same across fields. Furthermore, although the lognormal is sometimes called a thick-tailed distribution, all of the moments of the lognormal exist and the central limit theorem applies to it. Again, assuming that only the parameter \( \mu \) changes from field to field, one can normalize by the median rather than the mean. One has then normalized the citation distribution to the common distribution Lognormal[0, \( \sigma \)] for each field. We also note that Leydesdorff and Opthof (2011) commented that “... the ‘mean’ is not a proper statistic for measuring differences among skewed distributions” and suggested using the median. In the lognormal model with a fixed \( \sigma \), both the median and mean are scale parameters.

The discrete lognormal may be a candidate for citations to articles to a single subject and year (Evans, Hopkins, & Kaube 2012; Radicchi, Fortunato, & Castellano, 2008; Thelwall & Wilson, 2014), but it is not clear that \( \sigma \) is fixed across fields. The hooked power is another candidate (Thelwall & Wilson, 2014). But power laws are also candidates (Katz 2016; Yao et al. 2014) and it may depend on the time period. With such variation of opinion and variation of citation distributions, following
principle six of the Leiden Manifesto (Hicks et al. 2015), “Account for variation by field in publication and citation practices” seems a challenge for normalization by division or other formulas.

**Percentiles**

Knowing the full information of percentiles for two continuous distributions, allows one to transform them into a common distribution. Discrete distributions may pose problems. Eugene Garfield concluded his book (1979 p. 249) commenting on the power of percentile comparisons:

*Evaluation studies using citation data must be very sensitive to all divisions, both subtle and gross, between areas of research; and when they are found, the study must properly compensate for disparities in citation potential. This can be done very simply. Instead of directly comparing the citation count of, say, a mathematician against that of a biochemist, both should be ranked with their peers, and the comparison should be made between rankings. Using this method, a mathematician who ranked in the 70 percentile group of mathematicians would have an edge over a biochemist who ranked in the 40 percentile group of biochemists, even if the biochemist’s citation count was higher.*

In other words, the proper way to make comparisons would be by percentiles. But this method has its challenges too. It still requires an understanding and integrity of the field distributions. As Garfield (1979 p. 249) continued, “We still know very little about how sociological factors affect citation rates. There is still much uncertainty about all the possible reasons for low citation rates. And there is still much to learn about the variations in citation patterns from field to field.” It is still true today and poses a serious issues if citation analysis is to respect the first principle of the Leiden Manifesto (Hicks et al. 2015), “Quantitative evaluation should support qualitative, expert assessment.” We take the example of one field in which to perform citation analysis and normalization.

**Mathematics**

Details by subfield classification in mathematics was examined by Smolinsky and Lercher (2012). Even in a single Journal Citation Report field, the distribution of citations in subfields may be very different and at odds with expert assessment. They examined young researches who were the highest respected by expert assessment, i.e. Sloan fellows. They classified Sloan fellows using a classification by the joint committee of the mathematics, applied mathematics, and statistics societies. The researchers fall into 9 fields (the field of statistics was not included).

In terms of the expert assessment, Algebra seemed to be the highest assessed. It had by far the most Sloan fellows (32 of 99). Hiring of new PhD’s at the top 48 “Group I” mathematics departments was also greatest in algebra. However, this was in contrast to the bibliometric data. The algebra Sloan fellows were 7th out of 9 in the number of publications per fellow and 6th out of 9 for citations per paper.

Putting aside the contradiction between the two assessments, the variation in a single field among subfields is well known. Glänzel and Schubert (2003 p. 357) sum it up: “The classification of scientific literature into appropriate subject fields is…one of the basic preconditions of valid scientometric analyses. Publication activity and citation habits considerably differ among subfields.”

**Acknowledgments**

The author is grateful to George Cochran for a discussion on scale parameters in probability distributions.

**References**


Predicting the Future Impact of Scholars through Time-aware Academic Networks

Jun Zhang, Wenjie Kang, Jiaying Liu, Teng Guo, Ivan Lee, Feng Xia

f.xia@ieee.org
School of Software, Dalian University of Technology, 116620, Dalian (China)

Introduction
The scientific impact of scholars has been studied by researchers with diverse backgrounds for a long time. Predicting the scientific impact of scholars through utilizing the scholarly big data (Xia et al., 2017) can shed light on many problems, such as discovering potential collaborators, tracking research trends, and providing basis for awards or foundation applications etc. Therefore, predicting scholars’ scientific impact is of great significance and has drawn increasing interests.

Generally, predicting the impact of scholars can be divided into two main categories: citation-based metrics and network-based methods. Citation-based methods mainly consider the dynamics and distinctions of citations to predict the future influence of scholars (Sinatra et al., 2016). In network-based metrics, the different academic network topologies and various importance ranking algorithms are utilized to predict the impact of scholars. However, one important fact is neglected, i.e., the dynamics of the academic networks. Previous studies have employed different academic network structures to depict the real academic networks. As a matter of fact, academic networks evolve with time, for instance, articles will get new citations and scholars will cooperate with different authors. It is challenging to capture the dynamics of academic networks for better predicting scholars’ future influence. Motivated by the above mentioned phenomena, we propose the TARank method to predict scholars’ impact. Our method not only considers the mutual influence among different scholarly entities, but also captures the dynamics of heterogeneous academic networks to predict the future impact of scholars. Furthermore, the performance of our methods is explored on Microsoft Academic Graph (MAG) dataset.

TARank Method
As shown in Figure 1, our TARank method mainly consists of four main parts. We first construct four academic networks, i.e., time-aware paper citation network, time-aware coauthor network, time-aware paper-venue network, and time-aware paper-author network. The detail calculation procedures through utilizing these time-aware networks are as follows.

**Time-aware paper citation network**
There exists an edge between papers if one paper cites another in the time-aware paper citation network. Through analyzing the MAG dataset, we found that most recently published papers get more citations in the future comparing to papers that have been published for a long time. According to this phenomenon, we calculate the impact of articles through the time-aware weighted citation relationship between papers, and the calculation formula can be represented as:

\[
PR(p_i) = \frac{(1-d)N_p + d \sum_{j=1}^{m} \frac{PR(p_j)}{L(p_j)} \times e^{-\alpha(T_{co} - T_{ci})}}{1}
\]

where \(PR(p_i)\) represents the impact score of paper \(p_i\), \(m\) is the total number of papers that cite \(p_i\), \(L(p_j)\) is the sum of outgoing links of \(p_j\), and \(N_p\) is the total number of papers. \(T_{ci}\) is the current time, \(T_{ci}\) is the time when \(p_j\) cites \(p_i\), and \(d\) is the damping factor (set as 0.85).

**Time-aware coauthor network**
There exists an edge between authors if they have cooperated in the same paper in the time-aware paper citation network. The abilities of coauthors can significantly influence the qualities of the corresponding papers, and furthermore influence the impact of scholars themselves. Therefore, we assume that if a scholar cooperates with influential scholars recently, then he or she will more likely collaborate with them in the future. We calculate the impact of authors through the time-aware weighted cooperation relationship between authors, and the calculation formula can be represented as:

\[
PR(a_w) = \frac{(1-d)N_a + d \sum_{k=1}^{n} \frac{PR(a_k)}{L(a_k)} \times e^{-\alpha(T_{co} - T_{cw})}}{1}
\]

Figure 1. Architecture of TARank.
where $PR(a_w)$ represents the impact score of author $a_w$, $a_k$ is the cooperator of $a_w$, $m$ is the total number of $a_k$, $L(a_k)$ is the sum of outgoing links of $a_k$, and $N_a$ is the total number of authors. $T_{cp}$ is the current time, and $T_{cp}$ is the time when $a_k$ cooperates with $a_w$.

**Time-aware paper-venue network**

The time-aware paper-venue network contains two kinds of nodes, which are papers and their corresponding venues. One type of relationship is included, which is the publication relationship between papers and venues. The qualities of papers and venues influence each other, and papers that publish in recent years are more indicative of the future impact of venues. The time-weighted publishing relationship between papers and venues can be denoted as:

$$TW_{pq} = \beta \times e^{r-\gamma(T_{pq}-T_{pb})} - \theta$$  \hspace{1cm} (3)

where $TW_{pq}$ represents the weight of the edge, and $T_{pb}$ is the publication time of paper $p$. In order to capture the mutual influence among papers and venues, we set up specific updating procedures. We first apply the HITS algorithm to calculate the impact values of venues according to papers’ impact scores (Eq. (1)). Then according to the impact values of venues we get, the scores of papers can be updated.

**Time-aware paper-author network**

The time-aware paper-author network contains two kinds of nodes, which are papers and their corresponding authors. One type of relationship is included, which is the writing relationship between papers and authors. It is universally acknowledged that the impact of scholars can be conveyed through their papers, and on the contrary, the qualities of papers can reflect the authors’ capacities. Therefore, we assume that the recently published papers can more indicative of scholars’ future influence. The time-weighted writing relationship between papers and scholars can be denoted as Eq. (3). The values of authors (Eq. (2)) and papers according to the last step are used as initial values. We first update the impact score of scholars according to their papers’ values, and then update the impact value of scholars through their papers through the random walk process. Repeat the above two steps until convergence, i.e., the sum of the differences of scholars’ values computed at two successive iterations falls below 0.000000001.

**Experiments & Results**

The sub-dataset used for our experiments is acquired from the MAG dataset. It contains the detailed paper information of each scholar, which includes article's title, keywords, coauthors, the date of publication, venues, authors' affiliations and its citation relationships. To conduct our experiments, we choose scholars from computer science area and they have no missing information. We use the data before 2010 to get the results of each method, and take the citation counts of each scholar in the recent 5 years as the ground truth to verify the prediction performance.

We compare our TARank with the following methods to evaluate its effectiveness on predicting scholars’ impact. **TARank_notime** applies our algorithm on the traditional academic networks without time-aware functions. **MRCoRank** is introduced in (Wang et al., 2016). As shown in Figure 2, the average citation counts of the top ranking scholars of our TARank is the highest compared to other methods. The results demonstrate that our method can efficiently predict the impact of top ranking scholars.

**Conclusion**

In this paper, we propose the TARank method to predict the scientific impact of scholars, and the experiments on real dataset indicate that our method can efficiently predict the future impact of top ranking scholars. In future work, we will investigate the performance of our method from more aspects on more dataset.

**References**


Using Readership to Measure Interdisciplinarity

Chengzhi Zhang¹,² and Shurui Xu¹

¹. Department of Information Management, Nanjing University of Science and Technology, Nanjing, China, 210094
². Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University), Fuzhou, China, 350108

zhangcz@njust.edu.cn, 115107000840@njust.edu.cn

Introduction

Interdisciplinary research is an important topic in the library and information science. There are a lot of interdisciplinary measurement indexes, such as Rao-Stirling diversity (Stirling, 2007), Integration Score (Porter & Cohen et al., 2007), etc. These indexes are mostly based on co-occurrence of authors, institutions or references. These objects belong to internal information of papers, e.g. references data. However, few scholars measure the degree of interdisciplinarity according to papers’ external information, e.g. readership data. A reader, as a recipient of paper knowledge, belongs to a certain discipline. We can detect interdisciplinary knowledge output of papers according to their readership data. This paper proposes a new approach for interdisciplinary measurement which is based on the proportion of disciplines covered by readerships.

Methodology

Framework

Since reference and readership data is different in discipline classification system, disciplines in these two classification system must be unified. After unifying of disciplines classification system, we compute interdisciplinarity according to reference and readership data respectively. Correlation analysis between these two interdisciplinary measurements is presented. Finally, readership data is used to analysis trend of interdisciplinary development. Flowchart of interdisciplinary measurement analysis is shown in Figure 1.

Data

We collected readership data of papers from Mendeley¹ on June 14, 2016. As shown in Figure 2, ‘Readership Statistics’ is presented in every paper page. For instance, readership of paper “How far does scientific community look back?” included “34% Social Sciences”, “32% Computer Sciences” and “9% Arts and Humanities”.

The papers we chose were published on PLOS ONE² from 2012 to 2015. The reference data of corresponding papers was obtained from PLOS ONE xml files directly. The quantity distribution of papers is shown in Table 1. A total of 48,930 papers with 2,363,861 references were obtained. All these papers have readership data. There are about 63% papers whose corresponding discipline can be found in the unified disciplines classification system.

<table>
<thead>
<tr>
<th>Year</th>
<th>#Papers</th>
<th>#References</th>
<th>%Available References</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>10,871</td>
<td>525,242</td>
<td>65.91%</td>
</tr>
<tr>
<td>2013</td>
<td>14,580</td>
<td>704,291</td>
<td>64.37%</td>
</tr>
<tr>
<td>2014</td>
<td>13,322</td>
<td>637,937</td>
<td>62.71%</td>
</tr>
<tr>
<td>2015</td>
<td>10,157</td>
<td>496,391</td>
<td>60.41%</td>
</tr>
<tr>
<td>Total</td>
<td>48,930</td>
<td>2,363,861</td>
<td>63.43%</td>
</tr>
</tbody>
</table>

Unifying of Disciplines Classification System

We unify disciplines of classification system for reference and readership data based on Science-Metrix system. Science-Metrix system can be regarded as three-level classification tree (Archambault & Beauschesne et al., 2011). In this paper, we only use the first and second level of Science-Metrix system to reduce difference between Science-Metrix system and discipline classification system.

¹ http://www.mendeley.com
² http://journals.plos.org/plosone/
Computing Interdisciplinarity

We adapted a metric known as interdisciplinary distance (IDD) (Bromham and Dinnage et al., 2016) to measure the degree of interdisciplinarity of a paper. Interdisciplinarity degree of each paper based on references and readership data is denoted as Ref_IDD and Read_IDD respectively. The IDD index was derived from Phylogenetic Species Evenness (Helmus & Bland et al., 2007). The IDD index takes into account the evolutionary relatedness of disciplines to each other, as well as their relative abundances. The metric is standardized so it falls between 0 and 1. The formal definition of IDD is:

\[
IDD = \frac{m_{\text{diag}[C]} M - M' C}{M' M - m_m^2}
\]

where \( M \) is a \( n \times 1 \) column vector containing the values of \( m_1, m_2, \ldots, m_n \), \( M' \) is the sum of \( m_i \) (\( \sum_{i=1}^{n} m_i \)), \( \text{diag}[C] \) is a \( n \times 1 \) column vector giving the main diagonal of \( C \), and prime (') denotes the transpose.

Experimental result analysis

Correlation Analysis

Table 2. Spearman correlations between Read_IDD and Ref_IDD (N=48930)

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Read_IDD</th>
<th>Ref_IDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read_IDD</td>
<td>1</td>
<td>-0.018*</td>
</tr>
<tr>
<td>Ref_IDD</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Correlation between the interdisciplinarity based on references (Ref_IDD) and the interdisciplinarity based on readership (Read_IDD) is analysed according to 48,930 papers. As table 2 shown, the Spearman correlation coefficient of Read_IDD and Ref_IDD is -0.018 (p=0), indicating that there is a negative correlation between them.

Trend of Read_IDD

We monitor development of the readership disciplinary data from 2012 to 2015. The results are shown in figure 3. According to figure 3, we can see that trend of interdisciplinarity based on readership is increasing over time (about 7% growth). This result is almost the same as the change in Integration Scores over time (Porter & Cohen et al., 2007).

Conclusions

This paper uses readership data of papers to measure their interdisciplinarity degree. There is a negative correlation between references-based and readership-based interdisciplinarity. The trend of interdisciplinarity based on readership is increasing. Readership also shows that science is indeed becoming more interdisciplinary.

Acknowledgements

This work is supported by Major Projects of National Social Science Fund (16ZAD224), Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (MJUKF201704) and Qing Lan Project.

References


Trial for the Data-oriented management using data-set of Equipment sharing in Hokkaido University

Maiku Abe¹  Shingo Ebata²  Hiroshi Amitsuka³  Hiromitsu Uehara⁴  Satomi Tajima⁵  Tomokazu Yoshizawa⁶

¹mike_abcris.hokudai.ac.jp  ²ebashin@cris.hokudai.ac.jp  ³amiimi@phys.sci.hokudai.ac.jp
⁴uehara@oeic.hokudai.ac.jp  ⁵tajima.s@cris.hokudai.ac.jp  ⁶tomokazu_yoshizawa@cris.hokudai.ac.jp

Hokkaido University, Kita 21, Nishi 10, Kita-ku Sapporo 001-0021, Japan

Introduction

Nowadays, it becomes increasingly important to gain efficiency and usability of equipment and assets in universities to maximize outcomes of research and education within the limited resource and budget. This is also ranked as a high-priority government policy dealing with the development of reform and functional enhancement of universities in Japan.

The Global Facility Center (GFC) provides a variety of services related with the use of more than 150 leading-edge machines (named “open facilities”) and high technologies accumulated in the 140 years’ history of Hokkaido University to support research and educational activities for not only on-campus but also off-campus researchers, students, and engineers (Figure 1). Total number of users is now about 35,000 per year.

As the result of equipment sharing activity in Hokkaido University, we have already stored the dataset of users, hours of equipment on active and so forth for over 10 years.

In this research, we try to show one proposal on the future strategy of the equipment sharing from short term to long term using the dataset we stored.

Logic Model for sustainable project management

Logic Model is one of the management tools to visualize concrete activities and evaluate indicators for final goals of the organization. In Euro-American countries, some manuals to utilize the Logic Model are prepared. In Japan, Logic Model has been used to develop a basic program for policy assessment activities. Logic Model is utilized in hierarchical PDCA cycle like Figure 2. In previous studies, the post-project evaluation or performance evaluation using benchmark analysis for the case of the road asset management are proposed. Using these concepts, the management of research infrastructure in universities is more advanced and more efficient in every hierarchical level; field, tactical, and strategic levels.

Methodology

In this research, we focus on the strategic level, “Number of users using the sharing equipment.” Since the number of users is changing as time proceeds, time-series analysis is adopted as a trend analysis or future forecast methodology.
forecasting and calculating evaluation indicators using these datasets on a trial basis.

Results and discussions

Figure 3 shows one of the time-series datasets for number of users (solid line) and forecasting results. Just by visualizing the stored data, some trends are qualitatively grasped. Dotted line, grey area, and light grey area show forecasting data, 75%, 95% confidential interval, respectively. Furthermore, upper, middle, and bottom figures show 1, 3, and 5-year forecasts, respectively.

The prediction period in the time-series analysis is generally defined as 10%-20% of the sample size. However, three confidential intervals of forecasting results in figure 3 is almost the same range. In this research, 3-year forecasting data is adopted as an equipment sharing strategy.

As a trial consideration focused on the future trend of the utilization situation, the correlative relationship among research organizations is calculated. Figure 4 shows relationships of research institution (RI) A, B, and C. In this case, there is no relation between RI trend A and B (a-1, a-2). On the other hand, there is a relation between RI trend A and C (b). It is hoped that these results are basic information for investment strategy to purchase new leading-edge equipment, repair sharing equipment and so forth.

Figure 3. Time-series data of users on one of the sharing equipment. Solid line, dotted line, Grey area, and light grey area show historic data, forecasting data, 75% confidential interval, and 95% confidential interval, respectively. Furthermore, upper, middle, and bottom figures show 1, 3, and 5-year forecasts, respectively.

Figure 4. Correlative relationship among research organizations.

References

Computing the Influence of Disciplinary Keywords Based on h-Index

Si Shen¹, Peng Wu², Dongbo Wang³
¹ sszcgfss@gmail.com
Nanjing University of Science and Technology (China)
² dragonwu99@vip.sina.com
Nanjing University of Science and Technology (China)
³ wangdongbo0102@gmail.com
Nanjing Agricultural University (China)

Introduction
The keywords, especially ones with high frequency, are often applied to mining research focus and disciplinary core knowledge. But the keywords are mainly computing by counting frequency and cooccurrence relation. This paper, based on WOS (1985-2016) data of Artificial Intelligence (AI), computing the h index of each keyword by using the formula proposed by Hirsch (2005). From the perspective of statistical modeling, Malesios (2015) explores the variations on the standard theoretical models for the h-index. Bar-Ilan and Levene (2015) give a novel method of computing the h-index for web pages (the hw-rank) based on a variant of the h-index. Murray et al. (2016) find that the H-index is a useful quantitative measure for comparing different diseases.

Introduction to data source and computing method and Analysis of the distribution
All keywords of AI were directly extracted based on the paper from WOS. The data source gained from WOS is introduced as the following table 1.

<table>
<thead>
<tr>
<th>Set</th>
<th>Results</th>
<th>Search criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2047 PN= (artificial intelligence) Refined by: Research field= (Sciences) And Document type= (Article) And Language= (English) Databases= SCI/SSC Timespan= 1985–2016</td>
<td></td>
</tr>
</tbody>
</table>

The computing method is shown in the following Figure 1 based on the algorithm proposed by Hirsch.

Fig. 1 Computing method of h value of keyword

The paper computes the h value of each keywords of AI based on the computing method. Firstly, the relationship of amount of keyword and citation times is computed based on WOS (1985-2016) data of Artificial Intelligence (AI).

<table>
<thead>
<tr>
<th>Citation times</th>
<th>Amount of keyword</th>
<th>ln (amount of keyword)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>12674</td>
<td>9.44730793</td>
</tr>
<tr>
<td>1</td>
<td>3139</td>
<td>8.10748812</td>
</tr>
<tr>
<td>2</td>
<td>1945</td>
<td>7.57307256</td>
</tr>
<tr>
<td>3</td>
<td>1479</td>
<td>7.29912463</td>
</tr>
<tr>
<td>4</td>
<td>1112</td>
<td>7.01391537</td>
</tr>
<tr>
<td>5</td>
<td>849</td>
<td>6.74409186</td>
</tr>
<tr>
<td>6</td>
<td>811</td>
<td>6.69826854</td>
</tr>
<tr>
<td>7</td>
<td>693</td>
<td>6.54102999</td>
</tr>
<tr>
<td>8</td>
<td>591</td>
<td>6.381816017</td>
</tr>
<tr>
<td>9</td>
<td>513</td>
<td>6.240275845</td>
</tr>
<tr>
<td>10</td>
<td>421</td>
<td>6.042612834</td>
</tr>
</tbody>
</table>

According to the detailed value, the power-law between citation times and amount of keyword is shown. In order to observe the distribution phenomenon obviously, all keywords are processed based on the logarithmic computation. A scatter diagram which consists of citation times and amount of keyword is drawn. Figure 2 gives the whole distribution.

Fig. 2 The Power-law distribution of keyword

X-axis shows the citation times of keywords and Y-axis is the logarithmic computation result of amount of keyword. Figure 2 presents a linear trend which is consistent with power-law distribution. This paper, on the basis of the relation between h index in information output processing, Zipf’s Law, and Lotka’s Law proposed by Egghe (2006), proves the existence of such relation between keyword’s h index, the frequency and citation of keyword of WOS (1985-2016) data of Artificial Intelligence (AI). The following formulas explains the relationship.

\[ f(j) = \frac{C}{j^\alpha} \]

\( C > 0, \ \alpha > 1 \). \( f(j) \) means the total frequency of keywords of WOS (1985-2016) data of Artificial Intelligence (AI) whose citation times is \( j \). Therefore, the total frequency is:

\[ T = \int f(j) dj = \int \frac{C}{j^\alpha} dj = \frac{C}{\alpha - 1} \]
Then, the total frequency of keywords of WOS(1985-2016) data of Artificial Intelligence (AI) whose citation times is more than \( n \) is:

\[
\int f(j) \, dj = \frac{C}{\alpha - 1} j^{\alpha - 1} = T^{\alpha - 1}
\]

Let \( T^{\alpha - 1} = n \), then \( h \) is the solution to the equation.

\[
h = T^{1/\alpha}
\]

Formula presents the relation between \( h \) index and the total frequency of keywords (T) in WOS(1985-2016) data of Artificial Intelligence (AI). By verifying such relation, a mathematical relation between them can be concluded: when \( \alpha > 2 \), the total times cited of all keywords of WOS(1985-2016) data of Artificial Intelligence (AI) are:

\[
A = \int f(j) \, dj = \int \frac{C}{j^{\alpha - 1}} = \frac{C}{\alpha - 2} \quad A = T^{\alpha - 1} \quad \frac{1}{\alpha - 2}
\]

Therefore, Formula could be turned into

\[
h = \left( \frac{\alpha - 2}{\alpha - 1} A \right)^{1/\alpha}
\]

So, the relation between \( h \) index and A times cited of keywords of WOS(1985-2016) data of Artificial Intelligence (AI) is:

\[
h = \left( \frac{\alpha - 2}{\alpha - 1} A \right)^{1/\alpha}
\]

Meanwhile, the average times cited of each keyword of WOS(1985-2016) data of Artificial Intelligence (AI) could also be reached:

\[
\mu = \frac{A}{T} = \frac{\alpha - 1}{\alpha - 2}
\]

**Keyword’s \( h \) index computing**

The keyword’s \( h \) index is computed based on the computing method of Fig.1. All of the keyword’s \( h \) index are derived from the papers of WOS(1985-2016) data of Artificial Intelligence (AI). Table 3 presents the top 10 keywords with high \( h \) index and arranges them in descending order.

### Table 3 Top 10 keywords of WOS(1985-2016) data of Artificial Intelligence in order of \( h \) index

<table>
<thead>
<tr>
<th>Keyword</th>
<th>( h )-index of keyword</th>
<th>Frequency of keyword</th>
<th>Citation times of keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>artificial intelligence</td>
<td>66.</td>
<td>2194.</td>
<td>23057.</td>
</tr>
<tr>
<td>fuzzy logic</td>
<td>24.</td>
<td>207.</td>
<td>1078.</td>
</tr>
<tr>
<td>export systems</td>
<td>22.</td>
<td>134.</td>
<td>1583.</td>
</tr>
<tr>
<td>machine learning</td>
<td>20.</td>
<td>159.</td>
<td>1457.</td>
</tr>
<tr>
<td>genetic algorithms</td>
<td>20.</td>
<td>160.</td>
<td>1638.</td>
</tr>
<tr>
<td>data mining</td>
<td>19.</td>
<td>177.</td>
<td>931.</td>
</tr>
<tr>
<td>case-based reasoning</td>
<td>16.</td>
<td>96.</td>
<td>769.</td>
</tr>
<tr>
<td>evolution</td>
<td>15.</td>
<td>95.</td>
<td>1147.</td>
</tr>
</tbody>
</table>

In order to show the differences among the three values of \( h \)-index, frequency and citation times, the figure 3 gives the visualization results of top ten keywords.

The keywords computed by the \( h \)-index in Tab.3 and Fig.3 balance the frequency and citation times of keywords in all papers from WOS. So the keywords are more rational and scientific in distribution.

**Conclusion**

The computing method of keyword’s \( h \) index is put forward according to the definition of \( h \) index. The relationship between \( h \) index and frequency of keywords is explored, and the verification formulas is deduced based on the keyword of WOS(1985-2016) data of Artificial Intelligence (AI). The keywords \( h \) index of WOS(1985-2016) data of Artificial Intelligence (AI) are computed and shown in table and figure.

**Acknowledgments**

The WOS(1985-2016) data of Artificial Intelligence (AI) is provided by the Jiangsu Key Laboratory of Data Engineering and Knowledge Service and the exploring relationship is discussed with the researchers of the department of computer science and technology of Nanjing University. We thank you for giving us the help and guidance.

**Reference**


Correlation between scientific production, energy production and funding on renewable energies at the global level

Rafael Aleixandre-Benavent 1; José Luis Aleixandre Tudó 2; Lourdes Castelló-Cogollos 3; José Luis Aleixandre Benavent 2

1 rafael.aleixandre@uv.es
INGENIO (CSIC-Universidad Politécnica de Valencia) & UISYS (CSIC-Universidad de Valencia). Plaza Cisneros 4, 46003 València (Spain)

2 José Luis Aleixandre Tudó
Instituto de Ingeniería de Alimentos para el Desarrollo (IIAD). Universidad Politécnica de Valencia, Spain.

3 Lourdes Castelló-Cogollos
Departament de Sociologia i Antropologia Social. Universitat de València (Spain)

2 José Luis Aleixandre Benavent

Introduction
The burning of fossil fuels is a major cause of global climate change. The development of renewable energies (RE) is thus a significant strategy to mitigate global climate change (Cong, 2014; Cong, 2013). Meeting energy demands by increasing renewable energies is one of the crucial approaches of the power industry towards a sustainable development and climate change prevention. For example, China’s electricity sector is aiming to reduce with the use of RE nearly 50% of its greenhouse gas emissions by 2020 (Steenhof & Fulton, 2007). The first steps for the promotion of RE appeared in the 1990s from international environmental treaties such as the UN Framework on Climate Change in Rio, Brazil (1992) and the Kyoto protocol (Japan, 1997) (Chien & Hu, 2008). On the other hand, RE are becoming a very important asset for the development of the society, since during 2015 a 5% employment increase in the renewable energy sector was observed, to a total of 8 million jobs (direct and indirect) (REN21, 2017).

This work thus aims to investigate a possible correlation between scientific production and impact, power production and funding on global renewable energies (2007-2016).

Methods
Records on publications, citations and funded papers were obtained from Web of Science Core Collection (WOS) platform from Clarivate Analytics. A bibliometric analysis was performed to determine indicators of scientific productivity, impact, power production and funding. Key words for the search strategy included: biomass, geothermal, hydroelectric, solar, ocean and wind, all of them combined with “energy*”. The generic term “renewable energ*” was also taken into account. The key words were truncated if necessary to obtain variants of the same term. The search was limited to original articles within the 2007-2016 decade.

Our hypothesis includes: a) a possible correlation between power production on RE -expressed in Gigawatt hour (GWh)- and scientific production -as observed in a number of scientific published and funded papers; b) a possible correlation between power production and impact of the research.

Data on power production of renewable energies were extracted from the International Energy Agency (www.iea.com).

Results
A number of 12,167 papers that received 185,410 citations were recovered. The evolution of papers and citations is shown in Figure 1 with one third of the papers published in the 2015 and 2016 years.

![Figure 1. Number of published papers and citations per year](image-url)
Table 1 shows the 15 countries with more than 300 published papers. United States is the leading country in papers and citations, with China as the country with the highest percentage of funded papers (80%), followed by South Corea (77%) and Spain (61%).

Table 1. Most productive countries on renewable energies and funded papers

<table>
<thead>
<tr>
<th>Country</th>
<th>NA</th>
<th>NC</th>
<th>NFA</th>
<th>%NFA</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>2,320</td>
<td>46,390</td>
<td>1,375</td>
<td>59%</td>
</tr>
<tr>
<td>China</td>
<td>1,629</td>
<td>20,043</td>
<td>1,307</td>
<td>80%</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>1,007</td>
<td>16,873</td>
<td>548</td>
<td>54%</td>
</tr>
<tr>
<td>Germany</td>
<td>730</td>
<td>11,627</td>
<td>334</td>
<td>46%</td>
</tr>
<tr>
<td>Spain</td>
<td>729</td>
<td>12,253</td>
<td>447</td>
<td>61%</td>
</tr>
<tr>
<td>India</td>
<td>571</td>
<td>7,465</td>
<td>167</td>
<td>29%</td>
</tr>
<tr>
<td>Canada</td>
<td>474</td>
<td>8,527</td>
<td>214</td>
<td>45%</td>
</tr>
<tr>
<td>Italy</td>
<td>452</td>
<td>7,256</td>
<td>168</td>
<td>37%</td>
</tr>
<tr>
<td>Turkey</td>
<td>445</td>
<td>6,543</td>
<td>92</td>
<td>21%</td>
</tr>
<tr>
<td>Australia</td>
<td>425</td>
<td>6,996</td>
<td>188</td>
<td>44%</td>
</tr>
<tr>
<td>South Corea</td>
<td>403</td>
<td>4,156</td>
<td>312</td>
<td>77%</td>
</tr>
<tr>
<td>France</td>
<td>366</td>
<td>5,971</td>
<td>167</td>
<td>46%</td>
</tr>
<tr>
<td>Denmark</td>
<td>357</td>
<td>10,240</td>
<td>174</td>
<td>49%</td>
</tr>
<tr>
<td>Iran</td>
<td>350</td>
<td>4,360</td>
<td>66</td>
<td>19%</td>
</tr>
<tr>
<td>Japan</td>
<td>321</td>
<td>3,739</td>
<td>142</td>
<td>44%</td>
</tr>
</tbody>
</table>

In order to answer the question of whether there is a relationship between energy production and the number of funded articles, Figure 2 was investigated. As can be observed, there are countries in which there is very little correlation, such as United States, which shows a large number of both funded and not funded articles with a low RE production. By contrast, China doubles US in energy production despite having fewer published and funded articles. Other countries with a good correlation between the energy production and the scientific production, including both funded and not funded, are Canada, Italy, Japan and India. Some countries in which the correlation was also positive include Germany, France and Turkey. It should be noted that the country with the least relationship is United Kingdom; country that funds more than 50% of the articles, but has a low energy production. The same can be said for Spain and South Corea as they share a high percentage of funded articles but a low energy production.

Conclusions

Articles and energetic production on renewable energies has increased in recent times. The correlation between energy production, scientific production and funding is uneven, with countries showing good correlation (China, India, Canada, Italy, France and Japan) while others not (United States, United Kingdom, Spain and Australia). It will be desirable to investigate the behaviour of the RE individually and by countries with the aim of identifying the reasons that explain these discrepancies.

Acknowledgments


References


Figure 2. Energy production on Renewable energies, published articles (NA) and funded papers (NFA)
Publication in arXiv: a current bibliometric analysis of preprint publication benefits and drawbacks in science and technology

Antonia Ferrer-Sapena¹; Rafael Aleixandre-Benavent²; Fernanda Peset¹; Enrique A. Sánchez-Pérez³.

¹ anfersa@upv.es & mfpesetm@upv.es
DCADHA. Universitat Politècnica de València. Spain

² rafael.aleixandre@uv.es
Ingenio (CSIC- Universitat Politècnica de València). UISYS (CSIC-Universitat de Valencia), Spain

³ easancpe@upv.es
Instituto Universitario de Matemática Pura y Aplicada. Universitat Politècnica de València. Spain

Introduction

Prepublication ensures a better chance in the diffusion of the papers, but is not easy to know how far this practice can really improve some bibliometric parameters of the authors, as the number of citations to their papers or the publication of their articles in journals with higher citation indexes. The repository of scientific manuscripts arXiv is nowadays an important tool for diffusion of research, mainly in Physics and Mathematics but, recently, also in other branches of science and technology (Moed, 2007; Davis & Fromerth, 2007). The aim of this work is to study whether or not pre-publication in arXiv benefits the authors or rather produces a dispersion of citations that takes away relevance to their publications.

Methods

We searched the references to papers published in arXiv in the Web of Science database (WoS) in December 2015 in the field Cited Work of Cited Reference Search. After fixing the right reference, Google Scholar was used to determine if the paper was already published in a regular journal. This gave us a set of 557 preprints as a working sample. Once the set of preprints was identified, several analysis were done: a) how many preprints in arXiv appeared in references of papers published in journals that are listed in WoS; b) Difference among the year of standard publication and the year in which the preprint was deposited in arXiv; c) Cites computation for papers that were published in a standard journal and cites in arXiv of the preprints. In case the journal does not provide the number of citations to its papers, Google Scholar was used instead registered in Google Scholar, and the difference among these quantities.

Results

A1. Publication ratio and publication delay

Table 1 shows the total number of deposited and published papers, respectively, for some scientific fields that have been chosen by the bigger number of papers appearing.

<table>
<thead>
<tr>
<th>Scientific Areas</th>
<th>Deposited</th>
<th>Published</th>
<th>Ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astrophysics</td>
<td>208</td>
<td>154</td>
<td>74</td>
</tr>
<tr>
<td>Computer Science</td>
<td>50</td>
<td>36</td>
<td>72</td>
</tr>
<tr>
<td>Condensed Matter</td>
<td>101</td>
<td>76</td>
<td>75.2</td>
</tr>
<tr>
<td>Mathematics, Statistic, Nonlinear Sc.</td>
<td>56</td>
<td>36</td>
<td>64.3</td>
</tr>
<tr>
<td>Physics</td>
<td>139</td>
<td>72</td>
<td>51.8</td>
</tr>
<tr>
<td>Quantitative Biology</td>
<td>7</td>
<td>3</td>
<td>42.9</td>
</tr>
<tr>
<td>Total</td>
<td>561</td>
<td>377</td>
<td>67.2</td>
</tr>
</tbody>
</table>

The result depends a lot on the subjects. In the grouped area of Mathematics, Statistics and Nonlinear Sciences, the ratio obtained is 64.3. It may indicate that mathematicians consider relevant to publish in a standard journals those manuscript that the community consider relevant enough to be cited in their papers. The opposite behavior can be observed regarding “Physics”. It should mean that, up to a point, some authors considers to deposit in arXiv enough for providing visibility to their research.

Other interesting bibliometric information that can be obtained concerns the mean of the time that is needed for publishing a paper after it is deposited in arXiv. It is supposed that the authors introduce the paper in arXiv when they finish the research work, so after this moment the delay can be directly interpreted as exclusively due to the publication process. The results are shown in Table 2 for the
grouped specialties. It must be remarked that the dispersion of the result is very big (high variance).

Table 2. Publication delay for grouped specialties

<table>
<thead>
<tr>
<th>Subject area</th>
<th>Publication Delay (years)</th>
<th>Publication Ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astrophysics</td>
<td>0,8</td>
<td>74</td>
</tr>
<tr>
<td>Computer Science</td>
<td>1,1</td>
<td>72</td>
</tr>
<tr>
<td>Condensed Matter</td>
<td>0,8</td>
<td>75,2</td>
</tr>
<tr>
<td>Physics</td>
<td>1,5</td>
<td>51,8</td>
</tr>
<tr>
<td>Mathematics</td>
<td>2,5</td>
<td>64,3</td>
</tr>
<tr>
<td>Quantitative Biology</td>
<td>2,0</td>
<td>42,9</td>
</tr>
</tbody>
</table>

A2. Elapsed time for publication versus impact of the papers

For all the manuscripts, we computed the citations in arXiv (provided by INSPIRE). Then we applied two different rules depending on if the paper was finally published or not:

a) If the preprint was finally published, we found the number of citations in the journal where it was published, or in Google Scholar instead in case the journal did not provide this information (figure 1).

b) If the paper was never published, then we used as external measure of citations the number given by Google Scholar (figure 2).

A weak positive correlation may be observed in both cases. There are a lot of papers without citations inside arXiv that have a big citation rate when they are published in a standard journal, or in Google Scholar.

Figure 2. Citations in arXiv (by documents in arxiv) versus Citations in Google Scholar

Conclusions

1. There are not fixed rules that allow us to understand the general behavior of researchers regarding pre-publication of their scientific results. It seems to depend strongly on the particular specialty that is considered, and no general tendency is observed.
2. It seems that there is a lack of general reasons regarding motivation of researchers to deposit their manuscripts in arXiv. Short publication delay sometimes appears together with big ratio of publication. Disciplines with a high publication delay may have small publication ratios, what would imply that researchers lose interest in publication once their results are communicated. Thus, we must claim again about the need of measuring impact of deposit of manuscripts via number of citations to really consider this way of scientific exchange.
3. It seems natural to think that in the near future also consideration of non peer reviewed articles will enter in the evaluation systems, since their impact can also be measured in terms for example of citations or downloads, but the current state of the art does not allow to consider them as publications.


References


Introduction
The scientometric and social network analysis of key words has great potential to expand valuable understanding of the advancement of the research in a specific journal as well as in cases of emerging fields such as new technologies and processes (Jarvelin & Vakkari, 1993; Leydesdorf & Welbers, 2011; Wang et al., 2012). The purpose of this work is to analyse the evolution and relations of bibliometrics with other Information Science topics, as reflected in the scientific literature published in the Information Science Journal of broad scope entitled Online Information Review.

Methods
Key words included in a set of 758 papers included in the Web of Science database from 2000 to 2014 were analysed. We conducted a subject analysis considering the key words assigned to papers containing almost one bibliometric term. A social network analysis was also conducted to identify the number of co-occurrences between key words (co-words). The Pajek software was used to create and graphically visualize the networks. A threshold of more than 2 co-occurrences was established.

Results
Table 1 includes the distribution of papers according to the assigned key words in three five-year periods. Impact is the most frequently assigned key word (n=31), followed by Social networks (n=24), Citation analysis (n=15), h-index (n=13), Indicators (n=13) and Rankings (n=13). Other key words that highlight with more than 10 published papers are Impact factor (n=80) and Cluster analysis (n=11). Most key words increased during the three analysed periods, highlighting h-index, Indicators and Webometrics. Key words with an annual decreasing tendency of frequency include Cluster analysis.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Impact</td>
<td>31</td>
<td>1</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>Social networks</td>
<td>24</td>
<td></td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>Citation analysis</td>
<td>15</td>
<td></td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>h-index</td>
<td>13</td>
<td></td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Indicators</td>
<td>13</td>
<td>1</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Rankings</td>
<td>13</td>
<td></td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>Impact factor</td>
<td>12</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Cluster analysis</td>
<td>11</td>
<td>2</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Bibliometrics</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Collaboration</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Metrics</td>
<td>5</td>
<td></td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Webometrics</td>
<td>5</td>
<td></td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

Figures 1 to 4 show the network of co-occurrence of key words (co-words) over the three considered periods and over the entire period. During the first period, 2000-2004 (figure 1), the network comprised 9 key words. Three key words occupied
a more central position and a prominent intermediary position: Search engines, World Wide Web and Science.

Figure 1. Network of co-words in the 2000-2004 period

During the second period (2005-2009) (figure 2), the network comprises 13 key words and World Wide Web and databases were the key words with the most centrality.

Figure 2. Network of co-words in the 2005-2009 period

During the third period (2010-2014) (figure 3), the network of co-words included 16 key words, and Science and Search engines were the key words with the most centrality and connections.

Figure 3. Network of co-words in the 2010-2014 period

Figure 4 shows the network of co-words during the entire period.

Figure 4. Network of co-words along the complete analyzed period (2000-2014)

Conclusions

Our results reveal that Bibliometrics, as represented by Online Information Review, is an evolving discipline that draws on literature from a relatively wide range of subjects to embrace new lines of research (Chen & Xiao, 2016).

As expected, “classical” key words related to information systems and words related to the Internet environment and bibliometrics predominate. In the last years, new domains that have emerged and play a major role in stimulating research are investigations related to scientific publications based on bibliometric approaches to research evaluation, such as Citation analysis, Impact factor, h-index, Social network analyses and Rankings.

Increasingly, impressive growth has occurred in topics related to new technologies, new databases and products, such as electronic commerce, social networks, Google Scholar, Scopus and open systems.

References


Green Patent Value Evaluation Based on Social Network Analysis

Ke-Chiun Chang¹  Huiling Ou¹  Yen-Jo Kiang²

¹kechiun@gmail.com
Wuhan University, Wuhan (China)

²kiangyenjo@ctbc.edu.tw
CTBC Financial Management College (Taiwan)

Abstract
With the rise of green technology, green patent value evaluation is critical to the enterprise. In order to explore the influencing factors of green patent value, we take the LED industry as an example. Then we use the number of patents cited as the proxy patent value, and divide the patent into high-cited patents and low-cited patents based on the H index of LED patents. We also introduce the social network analysis, trying to explore the influencing factors of patent value with network centrality and network position from a dynamic view. The results show that the social network indicators out-degree centrality and effect size have a significant positive impact on the patent value.

Conference Topic
Patent analysis

Introduction
As the increasing environmental pollution and resources shortage problems, many measures have been taken. One of these measures is developing green technologies. Green technologies are also known as environmental sound technology, which refer to technologies and products using more clean and energy-efficient raw materials. Green technologies reflect humanity's introspection about the consequences of damaging the ecological environment resulted by modern sciences and technologies. The concept of green technologies was first proposed by Braun e. and Wield d. in 1994 (E. Braun & Wield, 1994), then got the wide acceptance. According to patent statistical manual published by the organization of economic co-operation and development (OECD) in 2009, about 80 technologies are regarded as green technologies, and they can be divided into 7 aspects, including environmental management technologies, water-related adaption technologies, biodiversity protection and ecosystem health technologies, climate change mitigation technologies related to energy generation, transmission or distribution technologies, capture storage sequestration or disposal of greenhouse gases and climate change mitigation technologies related to transportation and buildings.

Based on the high value of green technologies, green patents also got rapid development. Except patents’ traditional characters, such as novelty, creativity and practicability, green patents have the unique character--environmental friendly, which means that green patents are the results of highlight for environmental benefits with the requirements of environmental protection in modern society. Now, many countries have provided the quick patent censor system for green patents. For example, in 2009, the UK launched the first green channel for patent applications, aiming to promote green technology innovation and gain a leading position in global competition through the rapid review of green patent technologies.

The research and development of green technologies and green patents have brought both opportunities and challenges to enterprises. On the one hand, the development of emerging green industries has brought huge potential market for enterprises, which can enlarge the sales of products and lead to more profits. On the other hand, it takes additional costs for enterprises to meet the national environmental standards proposed by the policy. Besides, more and more consumers will consider the environmental impacts of products as a significant selection
standard due to the environmental value of green patents, which will lead to increasing preferences for green patented products. So, enterprises must upgrade their existing equipment and invest more in R&D to product more environmentally friendly products in response to these changes.

In response to the environmental trends, patents applied in environmentally-related areas are increasing in recent years. In China, there were more than 200 million patents in 2015 according to the data from Intellectual Property Office. But it is well known that applying for a patent requires a great cost, so the importance of patent should not rely on the patents’ number, but also the patents’ quality. According to the statistical data from OECD, which said that 37.7% of the world's environmentally-related innovations were patented in China, but only 3.9% of the world's environmentally-related innovations were developed in China, which means that the utilization of patents is far from expectation compared with developed countries, and the value of these patents need to be improved. So, it is urgent for enterprises to realize the important of evaluating patents and use their limited resources to build their patent pools, which can maximum the cost-effectiveness when faced with more and more patents application.

Although the green patents are receiving more and more attention, there are only a few formers research study about the value of green patent, and they are mainly focus on the institutional level. Some researchers have made some progresses by setting up the establishment of evaluation indicators, building the model and doing empirical analysis in the area of patents’ value. Except using patents’ own attributes such as the number of patents cited forward and backward, the number of patent claims to measure a patent’s value, some researcher also discovered some professional and comprehensive indicators, like patent’s innovation and mature of technology as evaluation indexes. But, they only considered the patents’ own attributes separately and ignored the role of the patents played in the process of technology evolution. Usually, the importance of a patent’s value is showed by how important the role it plays in the technical knowledge flow process, meaning that if a patent pushes forward the whole technical knowledge evolution, it values higher compared with other patents without doubt.

As an important part of green patents, LED has the concept of green energy and is at the forefront of the global electronics industry. LED has characters such as low power consumption, long life, fast response and non-polluting. These characters make it highly important in semiconductor lighting field. Optoelectronic Technology represented by LED is the most promising industry technology in 21 century. So, studying the values of LED patent will contribute to the same researches in green technological fields and provide reference to the green patent examination system. In order to study the evaluation system, we divide LED patents into two kinds which are highly cited patents and low cited patent based on H index. We introduce Social Network Analysis method and we use the patent as the nodes of the network, and make the citation relationship between patents as the connection between nodes. We conclude that the social network indexes, out-degree centrality and effect size have positive effects of patent value.

We hope this study could help enterprises to do better in evaluating patents and construct better patent pools and establish competition strategies matches themselves. Besides this paper also provides new ideas to other green technology fields, which can be reference to the patent examination system for review patents more efficient.

**Literature review and research hypothesis**

**Social network analysis**

Social network analysis (SNA) is the method of studying the social relationships and structures, which mainly study the internal network between different social actors (Wasserman & Faust, 1687)
Social network analysis was used to analyse the relationships between people, and it originated in the 1930s. After the rapid development, combined with mathematics, statistics and computer science and other disciplines, social network analysis became an interdisciplinary analytical method.

Social network is comprised by nodes and ties in sociology. Nodes imply actors, which may be the individual, team, enterprise or other social organizations. Ties imply the relationships between nodes, and scholars always describe ties by matrix and graph. Matrix and graph are the manifestations and the research foundation of social network. The distribution and characteristics of the relations in the social network can be analysed by using the matrix, and the structure between the nodes and the overall properties of the network can be observed by the network graph which connect the actors.

Since Yoon and Park (2004) applied social network analysis to patent research, more and more scholars have begun to build patent networks for analysis. For example, Gress (2010) pointed out that the internal and external reference in patent citation network can reflect the technical dependency and related contributions between the patent groups. Lee, Su, and Wu (2010) studied the conductive polymer nanocomposite industry. They constructed a patented network diagram for this industry, and discussed the evolution of the technology by calculating and analyzing the relative distance and network location of the patent in the network. Kim and Song (2013) introduced the social network to analyse the litigation relationship between the 26 major companies in the smartphone industry. In this study, we focus on the LED industry, using the social network to analyse the impact factors of patent value.

**H-index**

The h-index was originally used to analyse the individual academic achievements of the scientist and was a new indicator proposed by Jorge E. Hirsch in 2005 (Hirsch, 2005). Hirsh, in his study of the contribution of physicists, defined the "h-index" as: In the N papers published by the scientists, the maximum number of papers is quoted at least h times.

Compared with the traditional evaluation index, h-index takes both the number and the quality of scientists into account at the same time. Therefore, the H index become one of the most commonly used evaluation indicators. T. Braun, Glänzel, and Schubert (2006) first applied the h-index to the field of journal evaluation. Mingers, Macri, and Petrovici (2012) used the h-index as a measure of the quality and contribution of journals. They studied 455 journal samples in the field of business and management and analysed the h-index with the impact factor and the cites per paper. The conclusion shows that the h-index is superior compared with the traditional bibliometric index. Molinari and Molinari (2008) found that the h-index can be used to assess the scientific results and contributions of research institutions.

As for patents, Guan and Gao (2009) explored the relationship between the patent h-index and patent counts, citation counts, the mean family size, the result showing that the patent h-index is an effective index to evaluate the technical importance and influence of the assignee. Since then, h-index is widely used to evaluate the patent value. This study will follow the h-index in the field of patent research, calculating the LED industry patent h value. According the h-index, the patent citation is divided into two regions: the patented cited number of times higher than the h value and the number of patent cited less than h value.

**Patent citations and patent value**

Patent citations are the mutual references between patents and patents or non-patents. According to the different citation orientation, patent citations can be divided to backward citations and forward citations. Backward citations mean the patent cite the existing patents, and it is a continuation of the previous art, while the forward citations mean the patent is cited by other patents, reflecting the implication to follow-up arts.
In the past research, although the forward citations and backward citations can reflect the information about the related patents, but the usage of these two indicators are different. Generally, backward citations information is used to study the trajectory of research and development, technological diffusion, knowledge flow, etc. Forward citations information is usually consider as the indicator of patent economic value, technological innovation and implication. Trajtenberg (1990) first used the number of patents cited for patent value analysis, then many theoretical and empirical studies have proved that the use of the indicators to assess the effectiveness of patent value subsequently. For example, Harhoff, Schererc, and Vopeld (2003) found that there is a significant positive correlation between the number of citations and the value of patents. So, we continued the existing literatures, using the number of patent citation to evaluate the patent value, and divided the patents to high-cited patents and low-cited patents according the H index which evaluated through patent cited number.

**Network centrality and patent value**

Network centrality is the focus of social network research, and it is an important structural location indicator. In sociological research, centrality is often used to measure the importance of the individual, the status of superiority and social prestige. (Freeman, 1979), for example, argues that the central place in the social network reveals the rights and status of an individual in a group, which means that the more the actors are at the center of the network, the greater their influences. By applying network centrality to the patent citation network, the patent can be used to measure the position of the patent in the whole network and to identify the higher patent value in the whole patent network.

**Degree centrality**

Degree centrality evaluates the positions or superiorities of nodes in the social network, reflecting the direct influence of actors in the network. In social network, if an actor has many direct ties with lots of actors, then the actor is at the local position, and it has widespread influence. In directed graph, the degree centrality can be divided into the In-degree centrality and Out-degree centrality according to the direction between the actors and other actors. In-degree centrality refers to the number of other actors in the network to receive a relationship issued by the other, while the out-degree centrality refers to the network of an actor to the number of other actors to issue a relationship.

As the patent citation network is ordered network, according the different citation orientation between patents, forward citation can be reflected as the out-degree centrality, while backward citation can be reflected as the in-degree centrality. According to the previous section, the number of patent be cited as the evaluated indicator is overlap with the in-degree centrality of patents, so we only consider the out-degree centrality, and ignore the in-degree centrality of patents.

As for patents, the out-degree centrality reflects the references and inheritances from other patents, and this indicator also shows the degree of dependence of this patent on other science and technology. In this rapid develop modern society, most of the patents are based on previous skills, and they are products developed and improved based on previous skills. From this perspective, if a patent cited more prior arts, it shows that the technological area of this patent is relatively mature, and the technology this patent contains is mainly improvement of the previous mature skills. Therefore, if a patent cited more previous patents, it can be said that it is based on a large number of comparatively mature previous technologies, which means that the patent can integrate the existing knowledge and technology better, and its covering technical field is more broader compared to other patents. Allison, Lemley, Moore, and Trunkey (2004) pointed out that the more valuable patents are usually cited more advanced technology, and
they are more easily cited by other patents at the same time. Thus, the higher value a patent has, the more patents it will cite, which means the higher out-degree centrality the patent has. Therefore, we present the first hypothesis:

\[ H1: \text{Compared to low-cited patents, high-cited patents have higher out-degree centrality.} \]

**Closeness centrality**

Closeness centrality is the sum of the shortest distance from a node to all other nodes in the network. Closeness centrality evaluates the distance between a node to other nodes, the distance more short, the more easily of the nodes reach to the other nodes. In social network, if an actor is connected to many other actors through a relatively short path, the actor has a high degree of nearness to the center, meaning the actor is in the center position of the network. Freeman (1979) said that if all the paths an actor connected to other nodes are short, meaning that the actor is close to many other actors. This actor is less dependence on other actors in the process of information transfer, while a non-core positions must transfer information through a lots of actors. In other words, the actors occupying the near-center position are more efficient in communicating information throughout the network than other actors because of their shortest path to other actors. Such actors also tend to act as key roles, which can reflect quickly in the process of resolving the problems. Thus, closeness centrality is the primary measure of the ability of the actor not to be controlled by other actors or the efficiency of the transmission of information between actors.

Similar to the degree centrality, the closeness centrality of directed graphs can be divided into in-closeness centrality and out-closeness centrality. The closeness centrality refers to the sum of the shortest distances required by other actors to contact or cite actors, and it is a passive approach to the actor. In contrast, out-closeness centrality refers to the sum of the shortest distance the actor wants to associate with or quote from other actors, reflecting the active approach of the actor. Likewise, this article deals only with the outward-near centrality of patents, the sum of the shortest distances required for a patent to be directly or indirectly cited to all other patents. When the outward orientation of the patent is close to the central degree, it means that the patent can efficiently absorb the information and knowledge contained in the previous patent in the citation network and acquire the technical knowledge or information in other fields. In contrast, the inventor of the patent may more effectively cite prior patents to develop and improve upon extensive reference to other patents and may therefore be of higher technical content than other patents, and its value is also higher than other patents. Thus, we propose hypothesis 2:

\[ H2: \text{Compared to low-cited patents, high-cited patents have higher out-closeness centrality.} \]

**Eigenvector centrality**

Eigenvector centrality is another standardized measurement method of actors’ centrality. Combining the centrality of the actors contacted with the specific actor in the whole network, eigenvector centrality measures the most core actors of the network, which means that the centrality of an actor not only decided by the number of the actors contacted with it, but also the quality of these actors.

As for patent network, eigenvector centrality weighs the most core patent for that this measurement index measure the patent and its contact patents’ centrality at the same time. In the citation network, if one patent has many relationships with many other high centrality patents, then the patent is important in the network. On the contrast, a patent may be at the
center of a local patent citation network but not the entire network if this patent is connected to a number of a number of non-central patents.

So, we can conduct actors that connect to an actor having high eigenvector centrality are always connected by many other actors. Thus, if a patent has high eigenvector centrality, those patents which cite or cited by this patent also have high centrality. So we can say that a patent has higher eigenvector, the patent is closer to the core position of the patent citation network. So, in the entire patent citation network, the patent can influence many patents by direct or indirect contact, that is to say that the greater the influence of such a patent, the higher the value. So, the hypothesis 3 as follows:

**H3: Compared to low-cited patents, high-cited patents have higher eigenvector centrality.**

**Network structure and patent value**

In addition to the centralities of specific nodes in the network, the network structure also plays an important role in social network analysis. Research has often divided the social network structure into open networks and closed networks. Open structure can be linked with different groups, and can control the flow of information through the intermediary position, so as to obtain more advantages. This view takes Burt's Structural Hole Theory as the main representative (Burt, 1992). In Burt’s theory, structural holes are some of the gaps in network relationships, that is, there is an indirect link rather than direct link between some actors in network relationships and other actors. Structural Hole theory holds that the structure which have gaps between the actions can contain less redundant information, so structural holes the efficiency of the entire network of information exchange can be improved. Besides, the advantages of this network structure also derive from the intermediary opportunities created by information exchange, that is, in an open network structure, the actors can establish contact with groups that are not associated with each other and obtain more resource advantage by controlling the information channel.

In contrast to open networks, closed network structure can facilitate the flow of information within the network and help foster the trust between internal actors, which is the positive effect of network closure. Specifically, the actor in this closed network structure has a high joint density, so the cohesion of the network will be very strong, and if there is abnormal information or abnormal behavior in the whole network, each actor can quickly reflect soon, and the abnormal information or behavior can be spread quickly and be the excluded of the network, so that all actors will benefit.

There are two metrics for measuring the degree of network closure: Network Constraints and Network Effect Size. The network constraint coefficient measures the degree of dependence of an actor on the other actors in the network and the amount of redundant information contained in the relationship. The effect size is used to measure the overall influence of the structure-hole actor, which is defined as the sum of the non-redundant links in the network structure. In this study, we use the effect size to measure the network structure of a specific node. The effective size larger, the more non-redundant links in the network. Therefore, we presents the following hypothesis 4:

**H4: Compared to low-cited patents, high-cited patents have lower network effects.**
Research method

Sample selection and data collection
This study uses the US patented LED patent data in the Thomson Innovation database as of May 2011, including three technical areas: LED epitaxial growth, chip fabrication and chip packaging technology. Searching the technical information of the above technical fields, 40,330 pieces of LED patents was picked up. After screening and removing the patents which was isolated, a total of 4650 patents was remained. The 4650 patents were then numbered and sorted in descending order according to their cited number of times until the number of a patent was greater than the number of citations of the patent. Finally, we subtract 1 from the serial number, and the answer was the patented h-index.

After the initial calculation of the LED patent, h-index is 105, that is, in the 4650 patents, there are 105 patents cited more than its serial number, such a patent is defined high-cited patent. Then we selected the related patents based on the principle of high-cited patent and low-cited patent ratio of 1: 2 for data screening. However, due to the number of patent applications and patent claim counts, there are 17 high-cited patents cannot accurately match the relevant low-cited patents. Such a non-matching high cited patent are removed, finally a total of 264 patents were included in this study, including 88 high-cited patents and 176 low-cited patents.

We divided patents into high-cited patent and low-cited patent by using the H-index of the patent, and the H-index is used as the proxy variable of patent value. In order to construct the patent citation network, we use the patent as the node and the patent citation as the link between the nodes, then we convert the filtered patent data into the matrix required by social network analysis, which there is the relationship between the two patents in the matrix Marked as 1, and there is no citation relationship between the two patents are marked as 0. Then we import the matrix into the social network research software Ucinet and calculate the patent network center and network patent location indicators, and Logistic regression analysis was performed.

Variable operational definition
This study selects the patent cited as a proxy variable of patent value. According to the above calculation of the patent H index, the LED domain patent citations are divided to the high cited patent and the low cited patents. According to existing research, compared to the low cited patent, the high cited patent is more valuable.

Degree Centrality. Degree centrality is the number of other nodes that are directly associated with a particular node i, and is often seen as the degree of the node directly gets the flow information in the network. Degree centrality includes the out-degree centrality and the in-degree centrality. We only uses the out-degree centrality to analyze the nodes. For a network diagram with n nodes, the out-degree centrality of node i is calculated as follows: (1):

\[ C_{DO}(n_i) = d_o(n_i) = \sum_{j=1}^{n} x_{ij} \]

\( x_{ij} \) is the direct relationship between node i and node j, whose value is 0 or 1.

Closeness Centrality. Closeness centrality is the sum of the shortest path distances from node i to other reachable nodes. In general, nodes with shorter paths tend to have a higher closeness centrality than the other nodes. Similar to the degree centrality, closeness centrality can be divided into out-closeness centrality and in-closeness centrality, and we only consider the out-closeness centrality. It is calculated as follows (2):

\[ C_e(n_i) = \left[ \sum_{j=1}^{n} d(n_i, n_j) \right]^{-1} \]

\( x_{ij} \) is the distance from the node i to node j.
Eigenvector Centrality. Eigenvector centrality is the central index of the node \( i \) after considering the centrality of other nodes connected to \( i \), and it measures the degree of influence of the other nodes on the node \( i \). The centrality is calculated as (3):

\[
Ce(i) = \lambda^{-1} \sum_{j=1}^{n} a_{ij} e_j
\]

\( d(n_i, n_j) \) is the connected relation between node \( i \) and its adjacent matrix \((i, j)\). If connected, it is equal to 1, otherwise equal to 0, \( \lambda \) is the eigenvalue of the adjacent matrix, and \( e_j \) is the eigenvector corresponding to each eigenvalue.

Effect Size. Effect size is often used to measure the node’s impact on the entire network, reflecting the degree of non-redundant information network. The effect size of node \( i \) is calculated as follows:

\[
ES = \sum_i \left( 1 - \sum_q p_{iq} m_{jq} \right), q \neq i, j
\]

where \( j \) represents all nodes connected to \( i \) and \( q \) represents the nodes except for \( i \) and \( j \); \( p_{iq} m_{jq} \) represents the redundancy between patent \( i \) and \( j \), and \( p_{iq} \) represents the proportion of the relationship input from patent \( i \) to \( q \).

We mainly use the patent characteristic index as the control variables, including Claim Counts and Inventor Counts. Claims counts refer to the extent to which the patentee requests protection when submitting a patent application. The greater the number of patent claims, the more technical areas covered by patents. The inventor counts of patents refers to the sum of the number of patented inventors. The number of inventors of patents not only reflect the complexity of the patent, but also reflect the degree of R & D expenditure. The more inventors of the patent, the higher complexity of the patent may be, so the value of the patent may be higher.

**Empirical analysis**

*Descriptive statistics*

Table 1 is the descriptive statistical analysis of the independent variables and control variables in the model. From descriptive statistics, the mean and standard deviation of Out-degree Centrality, Effect Size and Inventor Counts are large, and the mean and standard deviation of other variables is relatively small.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Max</th>
<th>Min</th>
<th>Mean</th>
<th>S.D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outdegree</td>
<td>134.000</td>
<td>0.000</td>
<td>14.840</td>
<td>23.066</td>
</tr>
<tr>
<td>OutCloseness</td>
<td>0.030</td>
<td>0.021</td>
<td>0.021</td>
<td>0.004</td>
</tr>
<tr>
<td>Eigenvector</td>
<td>0.171</td>
<td>0.000</td>
<td>0.015</td>
<td>0.032</td>
</tr>
<tr>
<td>Effect Size</td>
<td>131.919</td>
<td>0.000</td>
<td>15.851</td>
<td>23.166</td>
</tr>
<tr>
<td>Claims Count</td>
<td>85.000</td>
<td>1.000</td>
<td>22.170</td>
<td>14.521</td>
</tr>
<tr>
<td>Inventor Count</td>
<td>13.000</td>
<td>1.000</td>
<td>3.000</td>
<td>1.956</td>
</tr>
</tbody>
</table>

**T-test**

Table 2 shows whether there is a significant difference between the average of the variables between high and low cited patents by using the T test. The results show that there are significant differences between Out-Degree Centrality, Out-Closeness Centrality, Eigenvector Centrality and Effect Size. Compared with the low-cited patent, the social network index of high-cited patent are significantly higher, and the disparities of the degree of out-degree
centrality and the effective size are most obvious. There is no significant difference in claims counts and inventor counts between two different types of patent.

Table 3 Different Characteristics between High-cited Patents and Low-cited patents

<table>
<thead>
<tr>
<th>Variables</th>
<th>highly cited patents</th>
<th>low cited patents</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out-degree</td>
<td>4.850</td>
<td>34.470</td>
<td>-12.264**</td>
</tr>
<tr>
<td>Out-Closeness</td>
<td>0.021</td>
<td>0.022</td>
<td>-5.011**</td>
</tr>
<tr>
<td>Eigenvector</td>
<td>0.004</td>
<td>0.037</td>
<td>-9.286**</td>
</tr>
<tr>
<td>Effect Size</td>
<td>6.040</td>
<td>35.135</td>
<td>-11.841**</td>
</tr>
<tr>
<td>Claims Count</td>
<td>21.540</td>
<td>23.420</td>
<td>-0.992</td>
</tr>
<tr>
<td>Inventor Count</td>
<td>3.010</td>
<td>2.990</td>
<td>0.067</td>
</tr>
</tbody>
</table>

Note: **p<0.05, *p<0.1

Results of Logistic regression analysis

According to the number of patent citations, the patent H index is used to divide the patent into the highly-cited patent and the low cited patent. These two kinds of patents are used as binary virtual dependent variables, the out-degree centrality, the out-closeness centrality, the eigenvector centrality and effect size as independent variables, the claim counts and inventor counts as control variables, then we perform the logistic regression analysis.

Table 4 is the regression analysis results. The results show that the out-degree centrality and the effect size of the patent network is significant, but the influences of these two index on the patent value are opposite. Among them, the out-degree centrality has a positive effect, while the effect size has a negative effect on the patent value. It is proved that hypothesis 1 and hypothesis 4 are established. On the other hand, the out-closeness centrality and eigenvector centrality does not have a significant effect on the patent value, proving hypothesis 2 and hypothesis 3 are not found.

Table 3 Results of Logistic regression analysis

<table>
<thead>
<tr>
<th>variables</th>
<th>model</th>
</tr>
</thead>
<tbody>
<tr>
<td>OutDegree</td>
<td>0.312**</td>
</tr>
<tr>
<td>OutCloseness</td>
<td>-314.801</td>
</tr>
<tr>
<td>Eigenvector</td>
<td>-10.113</td>
</tr>
<tr>
<td>EffSize</td>
<td>-0.186**</td>
</tr>
<tr>
<td>ClaimsCount</td>
<td>-0.001</td>
</tr>
<tr>
<td>InventorCount</td>
<td>-0.027</td>
</tr>
<tr>
<td>Intercept</td>
<td>4.801</td>
</tr>
<tr>
<td>-2 Log Likelihood</td>
<td>207.439</td>
</tr>
</tbody>
</table>

Note: **p<0.05, *p<0.1
Conclusions and discussion
Based on the number of patents cited in the field of LED technology, we divide the patents into high-cited patent and low-cited patent according H index, calculating the network centrality indexes by constructing the patent citation network to analysis the patent value. The results show that there are significant differences between the high-cited patent and the low-cited patent. Combined with the analysis of social network, it is found that there is a significant positive correlation between the out-degree centrality of patents and the patent value. The higher the degree of patent expulsion is, the higher the patent value is. On the contrary, effect size has significant negative impact on patent value. The larger the effective size of the patent network, the lower the patent value. Out-closeness centrality and eigenvector centrality does not constitute a significant impact on patent value.
Therefore, when assessing the value of green patents, the enterprise can construct the patent citation network and calculate the relevant social network index of the specific patent to determine the value of the patent and choose a relatively high value of the patent to build a favorable patent portfolio, which can achieve cost-maximizing of technological innovation and obtain sustained competition advantages. At the same time, enterprises can also build a citation network to analyze the history and development trends of technological fields, and identify the core technology to adjust their own innovation strategies continuously.
In the process of green patent applications, in addition to the characteristics of the patent itself, the enterprise should also place patents in the all patent network to consider its value. Due to the negative correlation between the effect size of the patent network and the patent value, the enterprise can apply green patent on the basis of extensive reference to the existing technology, and strengthen the technical cooperation with the same field for reducing redundant information of the entire network.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (71403191).

References


A scientometric method for assessing an institution’s scientific collaboration policy

Hamid Bouabid,
Faculty of Science, Mohammed V University in Rabat, 4, Avenue Ibn Battouta, BP 1014 RP, Rabat, Morocco, h.bouabid@hotmail.com

Abstract
The paper suggests a comprehensive method for an institution's scientific collaboration policy assessment. This method is based mainly on indicators: signed agreements, co-publications, their citations and scientific fields as front science. The method also uses a mapping of collaborative institutions and fields to assess the proximity of the institution in its collaborative cluster as well as multidisciplinarity in this cluster. Three major classes of collaborative institutions are derived from the method: (1) a class where there is coherence between co-publications and agreements; the university has to faithfully intensify its collaboration with this class; (2) a class where there is a substantial amount of co-publications but no collaboration agreement; the university should consider setting up a formal frame for collaborative research; and (3) a class grouping institutions which cited the university knowledge and considered as potential partners for the university.

Conference topic
Methods and techniques - Science policy and research assessment.

Introduction and background
Science and technology (S&T) are increasingly open, multidisciplinary and highly integrative. Therefore, collaboration offers several opportunities and contributes to increased productivity and excellence in S&T. Several authors have demonstrated that there is a strong correlation between scientific collaboration and research productivity and impact (Katz and Martin, 1997, Lee and Bozeman, 2005, He et al., 2009, Abramo et al., 2011a and Finlay et al., 2012). Defazio et al. (2009) modeled the effect of collaboration and funding on productivity for three periods: pre-funding, during funding and post-funding. They found that receiving funding increases researcher productivity by approximately 14%, while collaborating with a partner in the network in the post-funding period increases productivity by approximately 70%. The positive
effect of collaborations on the impact of papers has also been proven using citations or impact (Katz and Hicks, 1997, Guerrero-Bote et al., 2013, He et al., 2009, Levitt and Thelwell, 2010). While there exists a rich literature offering findings and conclusions on collaboration assessments at the scientist-micro-level (Traore and Landry, 1997; Melin, 2000; Hara et al., 2003; Jha and Welch, 2010; Lee and Bozeman, 2005), or at the regional, national or international-macro-level (Tijis and Glanzel, 2010; Hoekman et al, 2010; Choi, 2012, Bordons et al., 2013), there are far fewer assessments of scientific collaboration at the institutional level, as done by Adams (2005), Ponds et al. (2007), Ortega and Aguillo (2013), Barth et al. (2014), Han et al. (2014) and Yan and Guns (2015). Furthermore, the methods applied in these papers are different and less methodologically organized to link collaborative inputs (agreements) to outputs (co-publications) for policy issues. Indeed, for example, Ponds et al. (2007) make use of the gravity model to examine the proximity effect of collaboration between different kinds or similar institutions. The work by Barth et al. (2014) was based on network and cluster analysis (carried out using VOSviewer software) to identify similar institutions for collaboration and potential partner’s ones. Finally, Yan and Guns (2015) applied eight algorithms to collaboration at the author, institution and country levels for predicting and shaping the dynamics of collaboration starting from existing topology.

Data and Method

In this assessment, all frameworks, agreements, conventions and Memorandum of Understanding (MoU), were considered acts of collaboration. Other types of contracts, tenders, and sponsoring agreements or similar were excluded. During the whole period, UM5S signed 248 agreements.

The University Mohammed V - Souissi (UM5S) comprises the following 10 institutions:

- Medicine and pharmacy faculty;
- Dental medicine faculty;
- National high school of technology teaching;
- National high school of computer science and systems analysis;
- Law and Economics faculty in Souissi;
- Law and Economics faculty in Sale;
- Education sciences faculty;
- Institute for African studies;
- Institute for Arabic studies and research;
- University Institute for scientific research.

The Web of Science database was used to retrieve publications and co-publications (co-authored ones), where all kinds of publications were considered (article, review, proceeding, letter, editorial, etc). In order to compare agreements and co-publications, the analysis covers
the second period from 2010 to 2013 (end of June 2013). UM5S published almost 600 publications during this period, of which 63% were co-publications. In this analysis, a co-publication is defined as a paper authored by the UM5S and at least one other institution (national or international).

To map and analyze co-publications, Salton's Cosine (Salton, 1983) is used as the index for calculating the proximity between the entities (institution, field, key-words, etc) of the network. Figure 1 reports the proposed method to assess the policy of university scientific collaboration.

![Figure 1: Scheme and stream of the proposed method for collaboration policy assessment](image)

Figure 1 plots the method process. The first stage consists of compiling data and information as a policy inputs. The second stage refers to the process of the analysis itself comprising three levels: (i) the stock of active collaboration agreements, which represents, for a given year, the difference between the number of cumulative valid agreements and the number of expired agreements up to that year, (ii) the coherence between agreements and co-publications and (iii) the impact of co-publications as measured by their citations. To achieve this analysis five indicators are produced in order to build up the three classes of collaborative institutions. In another institution's assessment, this number will depend on its collaboration policy.

**Results and discussion**

For the whole period, the stock of active collaboration agreements has been calculated, as described in the Data and Method section. Its analysis is not part of this paper.
**Coherence between agreements and co-publications**

Through the indicator of co-publication, the method quantifies the share of each institution in the collaboration activities of the university. For UM5S, the method has shown a national co-publication dominance amongst overall university co-publications (Table 1). Major national collaborative institutions are the Specialized Hospitals in Rabat, Hassan II University in Casablanca, and Ibn Tofail University in Kenitra. The different institutions belonging to the CHU Ibn Sina are from now on not considered separately but included in it. Furthermore, Academy Hassan II of Science and Technology is excluded since it is more of a funding body and not a research institution. Similarly, and for coherence purposes, the Ministry of Health will not be considered at all as a collaborative institution.

On the international level it is also noted that there is a diversity of collaborative institutions (Table 1). It is worth mentioning that two-thirds of the 39 UM5S collaborative institutions (above the threshold of 10 co-publications) according to co-publications with UM5S are not under any collaboration agreement with UM5S, demonstrating a clear distortion between agreement as input proxy and co-publication as its output proxy. Cross-checking the results in Table 1 from either signed agreements and co-publications according to the method in Figure 1 allows us to build up two classes:

- **Class 1**: institutions with which the university has collaboration agreements as well as co-publications with their researchers, such as National Institute of Health, University Mohammed V Agdal, Ibn Sina University Hospital (CHU), University Ibn Tofail, CNRS-France, University Zaragoza, University Bordeaux, Université de Montréal, Université Laval, Cheikh Zaid Hospital, University Cadi Ayyad;

- **Class 2**: institutions with which there is no collaboration agreement, but there is substantial co-publication activity, such as: Military University Hospital Mohammed V, University Hassan II (Casablanca and Mohammedia), University Autonoma of Madrid, King Saud University, University Mohammed I, University of Technology Czestochowa, University Lodz, University of Kyoto, European Hospital Georges Pompidou, University Nancy 1, Hospital Virgen Camino, Medical Center Erasmus, University Valencia, University Rouen, CHU Limoges, University Liege, CHU of Tours, University Sidi Mohammed Ben Abdellah, University Paris 7.

While Autonoma University in Spain appeared to be a collaborative institution to the UM5 without any agreement, we found that the university had signed several agreements with other Spanish institutions but with no scientific output, such as: University Laguna, Universi
Holding, university Murcia, university Las Islas Baléares and University Cadiz. It is one of the illustrative cases of discrepancy between agreements and co-publications.

**Table 1: Breakdown of the UM5S's collaborative institutions (with a threshold of 10 co-publications)**

<table>
<thead>
<tr>
<th>Institution</th>
<th>Agreement</th>
<th>Number of Co-publications</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natl Inst Health</td>
<td>x</td>
<td>31</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Mohammed v agdal</td>
<td>x</td>
<td>31</td>
<td>Morocco</td>
</tr>
<tr>
<td>Ibn Sina University Hosp*</td>
<td>x</td>
<td>29</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Ibn Tofail</td>
<td>x</td>
<td>26</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Hassan II</td>
<td></td>
<td>26</td>
<td>Morocco</td>
</tr>
<tr>
<td>Military Hosp Univ Med V</td>
<td></td>
<td>25</td>
<td>Morocco</td>
</tr>
<tr>
<td>Hassan II Acad Sci &amp; Technol</td>
<td></td>
<td>22</td>
<td>Morocco</td>
</tr>
<tr>
<td>Hop Enfants Rabat*</td>
<td></td>
<td>19</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Autonoma Madrid</td>
<td></td>
<td>14</td>
<td>Spain</td>
</tr>
<tr>
<td>Hop Specialites *</td>
<td></td>
<td>14</td>
<td>Morocco</td>
</tr>
<tr>
<td>CNRS</td>
<td>x</td>
<td>14</td>
<td>France</td>
</tr>
<tr>
<td>King Saud University</td>
<td></td>
<td>14</td>
<td>Saudi Arabia</td>
</tr>
<tr>
<td>University Mohamed I</td>
<td></td>
<td>14</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Zaragoza</td>
<td>x</td>
<td>14</td>
<td>Spain</td>
</tr>
<tr>
<td>El Ayachi University Hosp*</td>
<td></td>
<td>13</td>
<td>Morocco</td>
</tr>
<tr>
<td>Czestochowa Technology University</td>
<td></td>
<td>12</td>
<td>Poland</td>
</tr>
<tr>
<td>University Lodz</td>
<td></td>
<td>12</td>
<td>Poland</td>
</tr>
<tr>
<td>Kyoto university</td>
<td></td>
<td>11</td>
<td>Japan</td>
</tr>
<tr>
<td>University Bordeaux</td>
<td>x**</td>
<td>11</td>
<td>France</td>
</tr>
<tr>
<td>University Montreal</td>
<td>x</td>
<td>11</td>
<td>Canada</td>
</tr>
<tr>
<td>University Laval</td>
<td>x</td>
<td>11</td>
<td>Canada</td>
</tr>
<tr>
<td>Minist Health</td>
<td>x</td>
<td>11</td>
<td>Morocco</td>
</tr>
<tr>
<td>Hop Cheikh Zaid</td>
<td>x</td>
<td>11</td>
<td>Morocco</td>
</tr>
<tr>
<td>Hop European Georges Pompidou</td>
<td></td>
<td>10</td>
<td>France</td>
</tr>
<tr>
<td>University Nancy I</td>
<td></td>
<td>10</td>
<td>France</td>
</tr>
<tr>
<td>Hosp Virgen Camino</td>
<td></td>
<td>10</td>
<td>Spain</td>
</tr>
<tr>
<td>Erasmus Med Center</td>
<td></td>
<td>10</td>
<td>Netherlands</td>
</tr>
<tr>
<td>University Valencia</td>
<td></td>
<td>10</td>
<td>Spain</td>
</tr>
<tr>
<td>University Rouen</td>
<td></td>
<td>10</td>
<td>France</td>
</tr>
<tr>
<td>CHU Limoges</td>
<td></td>
<td>10</td>
<td>France</td>
</tr>
<tr>
<td>University Liege</td>
<td></td>
<td>10</td>
<td>Belgium</td>
</tr>
<tr>
<td>Hop Specialities Rabat*</td>
<td></td>
<td>10</td>
<td>Morocco</td>
</tr>
<tr>
<td>CHU Tours</td>
<td></td>
<td>10</td>
<td>France</td>
</tr>
<tr>
<td>Inst Natl Oncol*</td>
<td></td>
<td>10</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Cadi Ayyad</td>
<td>x</td>
<td>10</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Sidi Mohamed Ben Abdellah</td>
<td></td>
<td>10</td>
<td>Morocco</td>
</tr>
<tr>
<td>University Paris 07</td>
<td></td>
<td>10</td>
<td>France</td>
</tr>
</tbody>
</table>

* Ibn Sina Hosp institutions
** including ENSEIRB, Univ Michel de Montaigne and univ Montesquieu
The same case holds for some French institutions. Indeed, while no agreement was signed with University Paris 7 despite many co-publications, we found that several agreements were set with institutions in Paris such as: University Paris 13, University Paris Dauphine, University Paris René Descartes, Institute for Languages and Civilizations.

**Impact of co-publications**

This part of the method is mainly based on science mapping which has been proven to be a useful tool in assessment.

1. **Impact of co-publications**

Collaboration with high-productivity partners tends to increase impact as proved by Katz and Hicks (1997) who found that while collaborating with an author from the home or domestic institution increases the average impact by approximately 0.75 citations, collaborating with an author from a foreign institution increases the impact by about 1.6 citations. Likewise, Abramo et al. (2011b) confirmed that more productive scientists and those with high impact collaborate more abroad than their colleagues (They warned, however, that the reverse was not always true). In this regard, we provide in Table 2 the impact indicator for institutions of the two classes in order to ascertain those with higher impact than that of UM5S. Table 2 shows that almost all of the collaborative institutions were more knowledge productive (not normalized to university size). Despite its slightly lower impact rate, but closer to that of UM5S, Ibn Tofail University is one of the small-size but performed universities in Morocco (Bouabid, 2014).

Three institutions were excluded for having a number of publications below the threshold of 100 publications during the same period: National Institute of Health (Class 1), Hospital Cheikh Zaid (Class 1) and Hospital Virgen Camino (Class 2). Two other institutions were excluded for having lower impact than that of UM5S (shown in grey in Table 2).

At the end of this stage of the method's stream, the two classes comprise the following:

- **Class 1**: University University Mohammed V-Agdal, University Ibn Tofail, CNRS-France, University Zaragoza, University Bordeaux, Université de Montréal, Université Laval, University Cadi Ayyad. The scientific collaboration with this class should be strengthened since there is correlation between co-publication practices and collaboration agreements;

- **Class 2**: University Hassan II (Casablanca and Mohammedia), University Autonoma of Madrid, King Saud University, University Mohammed I, University of Technology Czestochowa, University Lodz, University of Kyoto, European Hospital Georges Pompidou, University Nancy 1, Medical Center Erasmus, University Valencia, University
Rouen, University Liege, CHU of Tours, University Sidi Mohammed Ben Abdellah, University Paris 7. The scientific collaboration for this class is distorted and unbalanced and the university should consider setting up a formal framework for collaboration for its researchers to enhance their research activities.

Table 2: UM5S’s collaborative institutions with numbers of publications during the same period of study (with a threshold of 100 publications), citations and impact (citations/publication)

<table>
<thead>
<tr>
<th>Institution</th>
<th>Nb. Publications (p)</th>
<th>Nb. Citations (c)</th>
<th>Impact (c/p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Natl Inst Health</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Mohammed V-Agdal</td>
<td>799</td>
<td>4,477</td>
<td>5.60</td>
</tr>
<tr>
<td>Ibn Sina University Hosp</td>
<td>454</td>
<td>690</td>
<td>1.52</td>
</tr>
<tr>
<td>Ibn Tofail</td>
<td>328</td>
<td>592</td>
<td>1.80</td>
</tr>
<tr>
<td>CNRS</td>
<td>139,314</td>
<td>299,263</td>
<td>2.15</td>
</tr>
<tr>
<td>Zaragoza</td>
<td>7,396</td>
<td>38,321</td>
<td>5.18</td>
</tr>
<tr>
<td>Bordeaux</td>
<td>10,221</td>
<td>72,763</td>
<td>7.12</td>
</tr>
<tr>
<td>Montreal</td>
<td>21,626</td>
<td>139,397</td>
<td>6.45</td>
</tr>
<tr>
<td>Laval</td>
<td>11,461</td>
<td>70,534</td>
<td>6.15</td>
</tr>
<tr>
<td>Hosp Cheikh Zaid</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Cadi Ayyad</td>
<td>1,289</td>
<td>7,981</td>
<td>6.19</td>
</tr>
<tr>
<td>Military Hosp Univ Med V</td>
<td>235</td>
<td>211</td>
<td>0.90</td>
</tr>
<tr>
<td>Hassan II</td>
<td>773</td>
<td>7,237</td>
<td>9.36</td>
</tr>
<tr>
<td>Antonomia Madrid</td>
<td>9,284</td>
<td>67,624</td>
<td>7.28</td>
</tr>
<tr>
<td>King Saud</td>
<td>10,399</td>
<td>43,073</td>
<td>4.14</td>
</tr>
<tr>
<td>Mohammed I</td>
<td>492</td>
<td>6,654</td>
<td>13.52</td>
</tr>
<tr>
<td>Czestochowa</td>
<td>1,232</td>
<td>2,469</td>
<td>2.00</td>
</tr>
<tr>
<td>Lodz</td>
<td>7,056</td>
<td>27,089</td>
<td>3.84</td>
</tr>
<tr>
<td>Kyoto</td>
<td>28,153</td>
<td>170,073</td>
<td>6.04</td>
</tr>
<tr>
<td>Hop European Georges Pompidou</td>
<td>288</td>
<td>2,786</td>
<td>9.67</td>
</tr>
<tr>
<td>Nancy I</td>
<td>978</td>
<td>7,611</td>
<td>7.78</td>
</tr>
<tr>
<td>Hosp Virgen Camino</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Med Cntr Erasmus</td>
<td>20,636</td>
<td>162,426</td>
<td>7.87</td>
</tr>
<tr>
<td>Valencia</td>
<td>11,416</td>
<td>98,050</td>
<td>8.59</td>
</tr>
<tr>
<td>Rouen</td>
<td>2,704</td>
<td>18,644</td>
<td>6.89</td>
</tr>
<tr>
<td>CHU Limoges</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Liege</td>
<td>8,536</td>
<td>66,025</td>
<td>7.73</td>
</tr>
<tr>
<td>CHU Tours</td>
<td>482</td>
<td>2,940</td>
<td>6.10</td>
</tr>
<tr>
<td>Sidi Mohamed Ben Abdellah</td>
<td>309</td>
<td>2,635</td>
<td>8.53</td>
</tr>
<tr>
<td>Paris 7</td>
<td>16,099</td>
<td>163,552</td>
<td>10.16</td>
</tr>
<tr>
<td>UM5S</td>
<td>UM5S</td>
<td>596</td>
<td>1,085</td>
</tr>
<tr>
<td>Morocco</td>
<td>Morocco</td>
<td>8,542</td>
<td>25,053</td>
</tr>
</tbody>
</table>

*p*: is the number of publications during the period from January 1st 2010 to June 30th 2013.

**c**: is the number of citations gained by the publications *p* in the time window: January 1st 2010 to December 31st 2014.

-- : means that the number of publications is below the threshold
As a policy implication, for the Class 2 institutions a 'bottom-up' approach is required to identify university researchers involved who would contribute to setting up formal scientific collaboration with these institutions. In fact, the scientific collaborations were found to be mostly based on individual initiative as was reported by Bordons et al. (2013) rather than a result of policy action. Melin (2000) suggested that scientists themselves should choose with whom they would like to cooperate, and under which forms.

2. Multidisciplinarity in co-publications

To assess proximity to collaborative institutions and the multidisciplinarity in collaboration of UM5S, as a key component of the method clusters are drawn in a heterogeneous map crossing both collaborating institutions and scientific fields for the university in terms of co-publications (Figure 2). Despite its relative complexity, this kind of map provides a complete visualization and is of greater value by combining both institutions and fields. For the first point, higher proximity - e.g. closeness - is shown to be more prevalent with specialized institutions in Rabat such as Children’s Hospital (Hôpital des enfants), Ibn Sina University Hospital, Mohammed V Military Hospital and National Institute of Health, as well as with Ibn Tofail University (35 kilometers north of Rabat). For the second point, Figure 2 pictures a multidisciplinary collaboration with very limited institutions: Ibn Tofail University (Class 1) and Georges Pompidou European Hospital (Class 2) as identified by the overlap of more than two different circles - sub-clusters (marked with arrows in Fig. 2). Policy collaboration focus should be on these institutions as multidisciplinary collaborative ones.

3. Research Fronts/advanced knowledge in co-publications

In addition to the classification presented above, a qualitative indicator is employed to check the relevance of considering these classes in the university collaboration policy. The objective is to ensure that the topics addressed in the research collaboration with these institutions are in 'Research Fronts' as defined by Thomson Reuters in its Science Watch©. This part of analysis is not presented in this paper. However, it was found that the topics addressed in the scientific collaboration were generally in the Research Fronts such as: genetics, genomes and gene expression, tissue engineering, immunology, aging, thin films, remote sensing, neurosciences, polymer, etc.
Finally, considering the overall knowledge production stream, the results showed that some institutions from emerging countries such as China, India, Brazil and Turkey are likely to be potential collaborative partners for the university since they have been citing its papers (Table 3) and with respect to their fast-growing knowledge production. These institutions are grouped into a third class as potential partners. Surprisingly, no co-publication or agreement were found with any of these institutions. Only institutions having more than five citations each (as a threshold) to UM5S' publications are considered as citing institution from these countries.
Table 3: Breakdown by country's collaborative institutions of the number of signed agreements, co-publications and received citations (for the 15 first countries and in bold the first five scores)

<table>
<thead>
<tr>
<th>Country's collaborative institutions</th>
<th>Co-publications</th>
<th>Agreements</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Rank</td>
<td>Number</td>
</tr>
<tr>
<td>Morocco</td>
<td>743</td>
<td>1</td>
<td>118</td>
</tr>
<tr>
<td>France</td>
<td>114</td>
<td>2</td>
<td>52</td>
</tr>
<tr>
<td>Spain</td>
<td>33</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>USA</td>
<td>19</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Japan</td>
<td>19</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>17</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Canada</td>
<td>14</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Poland</td>
<td>14</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Germany</td>
<td>12</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Italy</td>
<td>11</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Belgium</td>
<td>7</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Netherlands</td>
<td>7</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Egypt</td>
<td>5</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>Ukraine</td>
<td>5</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>Portugal</td>
<td>3</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>China</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>England</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>India</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Brazil</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Turkey</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Setting this threshold, all Turkish and Indian institutions were excluded. After this filter, the UM5S's potential partners are listed in Table 4. Neither publications nor citations of the Chinese Academy of Science is reported, because it is more a funding and policy-making body than a research institution, even if its output is almost 120,940 publications during the period of study. Similar to the institutions of classes 1 and 2, all the institutions of class 3 were found to have a higher impact on the scientific community worldwide (between almost 4 and 6 citations per publication) compared to that of the UM5S (1.82).

Table 4: List of the UM5S's knowledge citing institutions from emerging countries

<table>
<thead>
<tr>
<th>Institution</th>
<th>Nb. Publications (p)</th>
<th>Nb. Citations (c)</th>
<th>Impact (c/p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tongji University (China)</td>
<td>12,217</td>
<td>52,242</td>
<td>4.28</td>
</tr>
<tr>
<td>Sichuan University (China)</td>
<td>16,316</td>
<td>95,782</td>
<td>5.87</td>
</tr>
<tr>
<td>Shanghai Jiao Tong University (China)</td>
<td>27,993</td>
<td>123,550</td>
<td>4.41</td>
</tr>
<tr>
<td>Universidade De Sao Paulo (Brazil)</td>
<td>42,022</td>
<td>230,347</td>
<td>5.48</td>
</tr>
<tr>
<td>Universidade Federal Do Ceara (Brazil)</td>
<td>3,872</td>
<td>14,678</td>
<td>3.79</td>
</tr>
</tbody>
</table>

At the end of the method's stream, the institutions of the three classes are listed in Table 5, each requiring a specific action but within the same scientific collaboration policy.
Table 5: List of the institutions for collaboration with UM5S according to the three classes.

<table>
<thead>
<tr>
<th>Class 1: Institutions having agreements and co-publications</th>
<th>Class 2: Institutions having co-publications without agreements</th>
<th>Class 3: Institutions citing the university’s knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mohammed V-Agdal</td>
<td>University Hassan II</td>
<td>Tongji University</td>
</tr>
<tr>
<td>Ibn Tofail university</td>
<td>Antonoma Madrid</td>
<td>Sichuan University</td>
</tr>
<tr>
<td>CNRS</td>
<td>King Saud University</td>
<td>Shanghai Jao Tong University</td>
</tr>
<tr>
<td>Zaragoza university</td>
<td>University Mohammed I</td>
<td>Universidade De Sao Paulo</td>
</tr>
<tr>
<td>Bordeaux university</td>
<td>University Czestochowa</td>
<td>Universidade Federal Do Ceara</td>
</tr>
<tr>
<td>Montreal university</td>
<td>University Lodz</td>
<td></td>
</tr>
<tr>
<td>Laval university</td>
<td>Kyoto University</td>
<td></td>
</tr>
<tr>
<td>Cadi Ayyad university</td>
<td>Hop Europeen Georges Pompidou</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nancy 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MC Erasmus (Rotterdam)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>University Valencia</td>
<td></td>
</tr>
<tr>
<td></td>
<td>University Rouen</td>
<td></td>
</tr>
<tr>
<td></td>
<td>University Liege</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CHU Tours</td>
<td></td>
</tr>
<tr>
<td></td>
<td>University Sidi Med Ben Abdellah</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Paris 7</td>
<td></td>
</tr>
</tbody>
</table>

4. Conclusion

A method is suggested for assessing a university scientific collaboration policy. Starting from data on agreements, co-publications, *Research Fronts* and citations, scientometric indicators were built along with science mapping in a comprehensive and organized method to inform about the coherence between co-publications as research collaboration outputs and collaboration agreements as a 'legal' frame input. The method stands to provide valuable inputs for policy assessment and design by identifying three distinctive institution classes for collaboration. The first class comprises institutions for which there were both co-publications and collaboration agreements; the university has to intensify its collaboration with this class. The second class comprises institutions with significant co-publications and higher impact but no collaboration agreement; the university should consider setting up a formal 'bottom-up' frame of collaboration to better enhance existing individual collaborating activities. The third class comprises institutions highly citing university knowledge and having higher impact, thus providing potential collaborative partners.

Acknowledgement

The author would like to greatly thank R. Bouabid for his kind help.
References


Bouabid H. (2014), Science and technology metrics for research policy evaluation: some insights from a Moroccan experience, Scientometrics, 101, 899-915.


Choi S. (2012), Core-periphery, new clusters, or rising stars?: international scientific collaboration among ‘advanced’ countries in the era of globalization, Scientometrics, 90, 25-41.

Defazio D., Lockett A., Wright M. (2009), Funding incentives, collaborative dynamics and scientific productivity: Evidence from the EU framework program, Research Policy 38 (2), 293-305.


Lee S. & Bozeman B., (2005), The impact of research collaboration on scientific productivity, Social Studies of Science, 35(1), 673-702.

Levitt J. M., Mike Thelwell M. (2010), Does the higher citation of collaborative research differ from region to region? A case study of Economics, Scientometrics, 85 (1), 171-183.


An Altmetrics Study of TOP100 Samples in 2016

Dan Zou¹ and Yi Han²
¹531479222@qq.com, ²hanyi72@swu.edu.cn
¹²College of Computer and Information Science, Southwest University(P R China)

Abstract
This paper takes the TOP100 literatures with the highest Altmetrics Scores in the Altmetric.com in 2016 as samples. Taking the advantage of SPSS19.0, the correlation analysis between Altmetrics score and citation counting was presented, and the Pearson correlation coefficient is 0.036, which means that the Altmetrics score does not correlate with the citation counting and the Altmetrics indicator maybe independent variable to assess the literature impact. Meanwhile, the multiple linear regression analysis was conducted, which the Altmetric scores was viewed as dependent variable and some component indicators as independent ones. The fitness between the independent variable and the dependent variable is very good, and the significance test of the variable coefficient is also verified.

Keywords
Altmetrics; Altmetrics scores; citation counting; correlation analysis; multiple linear regression analysis

Conference Topic
Altmetrics

Introduction
With the development of open access movement and social media, the impaction of scientific achievements is not only reflected in academic fields, but in social ones. Academic impacts reflect the academic status of researchers and the academic values in their discipline, and in some extents, social impacts are the cognitions perceived and understood by the ordinary people outside the academic peers and communities (Zhao Rongying et al, 2016). At present, more and more experts, scholars and non-academic communities began to exchange academic achievements through social media, and Altmetrics, reflecting the social impact in some extents, is born in such context. Altmetrics is an evaluation method based on social network, measuring the mentioned degrees by social media, to reflect the impaction of scientific achievements in a comprehensive and immediate way(Piwowar, 2013). Priem(2012) further noted that Altmetrics will be an important tool in next generation to measure the impaction of academic achievement.

In recent years, relying on its direct and rapid responding feedback mechanism, Altmetrics has been playing a more and more important role in the scientific evaluation of academic achievements. Many publishers and databases, such as Nature, Science, Elsevier, Wiley-Blackwell, NPG, and Springer, have begun to vigorously promote Altmetrics indicators to measure the concerning or concerned degrees for their published papers and works, meanwhile a large number of measuring tools have been developed to solve data collection and their computation, such as the typical
Altmetrics score for their recorded papers in Altmetrics.com.

For a long time, the citation counting was an important indicator for academic evaluation in science of science, which has been widely concerned by the author, the scientific management bureau and the relevant stakeholders. So what is the relationship between the Altmetrics indicators and citation counting? Many scholars have carried out a lot of related works. Thewall et al(2013) have found that there are strong relationship between citation counting and six Altmetrics indicators, including Twitter, and cited data from WoS (excluding the self-cited), but the relationship between the citation counting and the Google+ is very weak, and the other four Altmetrics indicators are very low. Wang et al(2016) pointed out that the correlation between PDF downloading index and citation counting is very strong, but the correlation between Altmetric score and citation counting is very weak, which is obviously lower than the Mendeley. Cabezas-Clavijo et al(2010) have been utilized 8945 articles in PLOS to verify the relationship between Scopus citation counting and scientific blog linkage numbers and tweets, which results are positive one between Scopus citation counting and scientific blog links and tweets. Eysenbanch(2011) selected 55 highly cited papers in JMIR and then found that 75% of selected papers were recommended by Altmetrics tools. LiXuemei et al.(2012) used 1613 articles published in Nature and Science in 2007 to validate the relationship between cited indicators (WoS and Google references) and the readers numbers in document management tools (Mendeley and CiteULike), and the conclusion showed that academic references have a strong correlation with readers numbers in Mendeley and CiteULike, where the correlation coefficient between WoS references and Mendeley readers is 0.55, and the one between Google Scholar references and readers numbers in Mendeley is 0.6. Torres et al.(2013) also found that there was a strong correlation between citation counting and reader number in document management tool.

As the above have showed, there are more and more studies to explore the correlation between the Altmetrics indicators and citation indicators, but these studies are focused on the relation between the citation number and single Altmetric indicator, and few studies focused on the correlation between citation counting and aggregate variable, such as Altmetric score. This paper attempts to identify the relationship between them, and to analyze the contribution of every indicator to the Altmetrics score at the same time.

Data and Methods

Data collection

In order to identify the relationship between Altmetric score and citation counting, the 100 Top academic papers (following called TOP100) in Altmetrics.com, ranked by the Altmetric score, were selected, and some related data, such as the data about Twitter, Blog, Facebook, Google+, were downloaded on January in 2017. Meanwhile, the corresponding citation data are retrieved from Web of Science.
Research methods

The linear trend was showed in the scattering plot between Altmetric score and citation counting, so the Pearson correlation analysis was utilised to address the relationship. As we know, the Pearson correlation is used to describe the degree of linear correlation between the variables, and its correlation coefficient is in the range between -1 and +1. The greater the correlation coefficient is, the stronger the correlation between variables. In fact, if the value is more than 0.8, the two variables can be viewed as equivalent or alternative ones.

In order to analyze the contribution of every indicator to the Altmetrics score, Altmetrics score of the TOP100 samples were regarded as the dependent variable, and the other indicators, such as Twitter, Blog, Facebook, Google+ score, were viewed as independent variables. Multiple linear regressions is a excellent method to explore the relationship between the multiple explanatory variables (independent variables) and the explained variable (dependent variable). So the Multiple linear regression was used to measure the contribution of every elements to the Altmetrics score.

Results

Correlation between Altmetrics Indicators and Citation Counting

The related data of TOP10 articles in the TOP100 sample were listed in table 1. From the table, we can see that Altmetrics score ranking has a certain difference to the WOS citation frequency. In the TOP10 articles, the highest-cited literature has been cited 128 times, but Altmetrics score was ranked in 6th; the first ranking literature in Altmetrics score is 8152, but its citation number only is 15. So we can see that non-highly cited literature may also have a relatively high Altmetrics scores and high Altmetrics scoring documents may not be highly cited. Focusing on the publication time, most of the high Altmetrics score literatures were published in 2016, so the less citation can be interpreted easily. In contrast to the papers published earlier, later papers in publishing time, if they focused on academic fronts, may get more attention immediately and have higher Altmetrics score in a minute. But citation indicators is the opposite, generally the citation is taken place after publishing one year or more. So the Altmetrics indicators have very fast responses and feedbacks on the latest hot results, which is just another way to solve the time-delayed problem in traditional citation indicator and impact factor (IF), and readers can grasp the latest front and hot researched in time (Li et al, 2016).

Table 1 Altmetrics score and the corresponding citation data of WOS (TOP10)

<table>
<thead>
<tr>
<th>No.</th>
<th>title</th>
<th>Altmetrics score</th>
<th>WOS cited</th>
<th>Published time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>United States Health Care Reform: Progress to Date and Next Steps</td>
<td>8152</td>
<td>15</td>
<td>2016.8</td>
</tr>
</tbody>
</table>

1712
Note: “Observation of Gravitational Waves from a Binary Black Hole Merger” has not been included in web of science.

Table 2 is the result of the correlation between Altmetric scores and citation numbers. The Pearson correlation coefficient is 0.036, and the two-tailed significant value Sig. (2-tailed) is 0.732. According to the theory of correlation analysis, the coefficient between the two variables is very weak, in some extents it can be considered there is no correlation between them, namely the two variables are dependent one. Thus, it can be considered to some extent that the Altmetrics score and the citation counting are not substitutive, but complementary and mutually relationship. The scores from Altmetric.com only reflect the quality and quantity of online attention to the papers, rather than reflect the quality of scientific research (Curry, 2014). Traditional indicators just offset the deficiency of Altmetrics ones, and Altmetrics indicators also compensates the shortcomings of traditional ones, which can’t highlight the social attentions.

<table>
<thead>
<tr>
<th></th>
<th>Medical error—the third leading cause of death in the US</th>
<th>5047</th>
<th>17</th>
<th>2016.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Evidence for a Distant Giant Planet in the Solar System</td>
<td>4372</td>
<td>34</td>
<td>2016.2</td>
</tr>
<tr>
<td>4</td>
<td>Sugar Industry and Coronary Heart Disease Research: A Historical Analysis of Internal Industry Documents</td>
<td>4429</td>
<td>3</td>
<td>2016.11</td>
</tr>
<tr>
<td>5</td>
<td>Zika Virus and Birth Defects—Reviewing the Evidence for Causality</td>
<td>3808</td>
<td>128</td>
<td>2016.5</td>
</tr>
<tr>
<td>6</td>
<td>The Association Between Income and Life Expectancy in the United States, 2001-2014</td>
<td>3841</td>
<td>41</td>
<td>2016.4</td>
</tr>
<tr>
<td>7</td>
<td>Effect of Wearable Technology Combined With a Lifestyle Intervention on Long-term Weight Loss***</td>
<td>3225</td>
<td>2</td>
<td>2016.9</td>
</tr>
<tr>
<td>8</td>
<td>Mastering the game of Go with deep neural networks and tree search</td>
<td>3130</td>
<td>93</td>
<td>2016.2</td>
</tr>
<tr>
<td>9</td>
<td>The new world atlas of artificial night sky brightness</td>
<td>3412</td>
<td>2</td>
<td>2016.6</td>
</tr>
<tr>
<td>10</td>
<td>Evidence for a limit to human lifespan</td>
<td>3066</td>
<td>1</td>
<td>2016.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Pearson Correlation</th>
<th>Cited Frequency</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>.036</td>
<td>.732</td>
</tr>
<tr>
<td>2</td>
<td>Sig. (2-tailed)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>93</td>
<td>93</td>
</tr>
<tr>
<td>4</td>
<td>Pearson Correlation</td>
<td>.036</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Sig. (2-tailed)</td>
<td>.732</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>N</td>
<td>93</td>
<td>93</td>
</tr>
</tbody>
</table>

Table 2 Correlations between Altmetrics score with Cited Frequency
Note: there are seven articles were not included in web of science in the retrieving time.

**Contribution of the components Indicators in Altmetrics Scores**

Altmetric.com has collected all of the online awareness data of papers from social networks (such as Twitter, Blog, Facebook, and Google+), document management systems (such as Medenley, CiteULike, and F1000), mainstream media (News Outlets) as well as multimedia (YouTube), and created a new comprehensive indicator, Altmetric score, which took the values to represent the social impact of the literature (Wang et al., 2014). Table 3 has showed their data for some components of the Altmetrics score (only the first 10 articles are listed). First, the data distributions are not balance, namely that the values of some components are relatively larger, and others are smaller. To the Altmetrics score, the greater the value of components, the more important it is, so they have played an important role in the comprehensive index. The News Outlets, the mentioned times of Twitter and Facebook, and the number of Mendeley readers have played important roles to reflect the social impact.

### Table 3 specific indicator score of TOP10 literature

<table>
<thead>
<tr>
<th>No.</th>
<th>score</th>
<th>News Outlets</th>
<th>blogs</th>
<th>twitter</th>
<th>facebook</th>
<th>Google+</th>
<th>Mendeley</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8152</td>
<td>215</td>
<td>40</td>
<td>8142</td>
<td>200</td>
<td>35</td>
<td>54</td>
</tr>
<tr>
<td>2</td>
<td>5047</td>
<td>358</td>
<td>35</td>
<td>2430</td>
<td>268</td>
<td>49</td>
<td>277</td>
</tr>
<tr>
<td>3</td>
<td>4372</td>
<td>371</td>
<td>58</td>
<td>1063</td>
<td>91</td>
<td>156</td>
<td>219</td>
</tr>
<tr>
<td>4</td>
<td>4429</td>
<td>308</td>
<td>40</td>
<td>2084</td>
<td>212</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>3808</td>
<td>330</td>
<td>28</td>
<td>1327</td>
<td>143</td>
<td>16</td>
<td>513</td>
</tr>
<tr>
<td>6</td>
<td>3841</td>
<td>272</td>
<td>41</td>
<td>1796</td>
<td>36</td>
<td>23</td>
<td>237</td>
</tr>
<tr>
<td>7</td>
<td>3225</td>
<td>292</td>
<td>21</td>
<td>1006</td>
<td>54</td>
<td>9</td>
<td>66</td>
</tr>
<tr>
<td>8</td>
<td>3130</td>
<td>156</td>
<td>46</td>
<td>1892</td>
<td>22</td>
<td>39</td>
<td>3091</td>
</tr>
<tr>
<td>9</td>
<td>3412</td>
<td>295</td>
<td>27</td>
<td>1275</td>
<td>92</td>
<td>45</td>
<td>42</td>
</tr>
<tr>
<td>10</td>
<td>3066</td>
<td>260</td>
<td>31</td>
<td>1201</td>
<td>15</td>
<td>18</td>
<td>236</td>
</tr>
</tbody>
</table>

In order to reflect the distribution of each component or variable, the average value of each variable are calculated and ranked. Table 4 has showed the calculated average value of Altmetrics score and the component indicators. From the point of view of magnitude, Tweeters, Mendeley readers, and News Outlets are the first group, Facebook users, Blogs, and Google+ users are the second one, and the remained are the third one.

### Table 4 the distribution of the average score of the indicators

<table>
<thead>
<tr>
<th>No.</th>
<th>indicator</th>
<th>average</th>
<th>No.</th>
<th>indicator</th>
<th>average</th>
<th>No.</th>
<th>indicator</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>score</td>
<td>2305.40</td>
<td>7</td>
<td>Google+ users</td>
<td>19.15</td>
<td>13</td>
<td>Q&amp;A thread</td>
<td>0.24</td>
</tr>
</tbody>
</table>
From the statistics can be found, Altmetric.com's TOP100 papers were mostly published in 2016. Altmetrics scores are ranging from 1619 to 8152, and average total score is 2305.40. From the average value of each indicator, social networking Twitter has made the greatest contribution, with an average of 942.49 tweeters mentioned or forwarded. The second is document management system Mendeley, which is saved and shared by 261.67 Mendeley users. The third is news outlets, which average number of individual papers reported by the media was 183.91 times. Facebook and blogs are also the main data source for Altmetric.com, and the average of each paper was 48.42 and 21.02 users discussed and mentioned. The data sources, such as book reviewers, policy source, peer review, Q&A thread, and research highlight platform are less.

To further illustrate the impact of each indicator on the Altmetrics score, multiple linear regression analysis was conducted, which the Altmetric scores was viewed as dependent variable and some component indicators as independent ones. Table 5 and Table 6 show the main results.

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.994</td>
<td>.987</td>
<td>.986</td>
<td>104.533</td>
</tr>
</tbody>
</table>

As can be seen in the results of goodness-of-fit test in Table 5, the Adjusted R Square is 0.986 and very close to 1, which means that the fitness between the independent variable and the dependent variable is very good. More than 98% of the dependent varieties can be explained by the independent variable, and the error caused by random factors is less than 2%. Based on these data, a conclusion can be drawn that the selected independent variable is more complete and the established regression
equation model is valid in a whole.

What does each independent variable devote to the dependent variable? The regression coefficient of each independent variable in Table 6 can give us the answers, which shows the change ratio of dependent variable caused by the unit variety of independent variables. Before examining the specific effects, we must first confirm the validity of variable coefficients. Except the two variables of Twitter and Facebook, which significant levels are slightly greater than 0.05, the significant level of other variables is far less than 0.05. So the coefficient validity of all the variables can be convinced. Two other test indexes, collinearity tolerance and statistical VIF, make us believe the mutual independence of all the interpreting variables.

Table 6 The variable coefficient of each indicators on Altmetrics score

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std.Error</td>
<td>Beta</td>
</tr>
<tr>
<td>1</td>
<td>(Constant) -65.495</td>
<td>37.509</td>
<td>-1.746</td>
</tr>
<tr>
<td></td>
<td>twitter  .723</td>
<td>.014</td>
<td>.841</td>
</tr>
<tr>
<td></td>
<td>Mendeley  .056</td>
<td>.030</td>
<td>.030</td>
</tr>
<tr>
<td></td>
<td>news     8.025</td>
<td>.180</td>
<td>.625</td>
</tr>
<tr>
<td></td>
<td>facebook .545</td>
<td>.279</td>
<td>.029</td>
</tr>
<tr>
<td></td>
<td>blogs    6.775</td>
<td>1.192</td>
<td>.086</td>
</tr>
<tr>
<td></td>
<td>Google+  1.586</td>
<td>.437</td>
<td>.054</td>
</tr>
</tbody>
</table>

a. Dependent Variable: score

The coefficients of all the independent variables are positive, and the variable coefficient of News is the largest one, indicating that it has the most obvious impact on the Altmetric score, which implies that once an article is mentioned by the mainstream news media, the Altmetric score correspondingly increased 8.025. The following variable is Blogs, which coefficient is 6.775, and the third one is Google+, which coefficient is 1.586. The impaction of other independent variables is relatively less.

Contrasting with the data of Table 4, we can found some interesting phenomena. Although the data provided by Twitter, Facebook, Mendeley etc. network tools is larger, the contribution rate to Altmetric scores is not significant in terms of single data impact, especially Mendeley, whose unit contribution is only 0.056. Thus, when the comprehensive metrics indicator, Altmetric scores, was created by Altmetric.com, the mainstream media and social networks have been given more higher weighty, and the document management systems relatively lower.

According to Table 6, we can get the multiple linear regression equation to Altmetrics total score:
\[ Y = -65.495 + 0.723 X_1 + 0.056 X_2 + 8.025 X_3 + 0.545 X_4 + 6.775 X_5 + 1.586 X_6 \]

Among them, \( Y \) stands for Altmetrics score, -65.495 for constant, \( X_1 \) for twitter, \( X_2 \) for Mendeley, \( X_3 \) for news, \( X_4 \) for facebook, \( X_5 \) for blogs, and \( X_6 \) for Google+. If the values of all independent variables are given, the Altmetrics score can be computed and the error can be predicted.

**Conclusions and Discussion**

Altmetrics indicators mainly focus on the readers’ number of academic literatures, as well as their interaction behavior, such as recommendation and comments, through online social media and their depth of social communication in traditional media to measure their social influences and social concerns, but less on involving academic influences related to the quality of articles. While the traditional citation indicators directly measure the academic influences through citation counting, which directly reflect the academic value and influences of scientific researches, but it cannot reflect the social influences. In our studies, the correlation between Altmetrics scores and citation counting give us good example. In terms of the literature evaluation, Altmetrics is only a complementary indicator rather than alternative in essence, so it is inevitable for us to combine the traditional indicators and Altmetrics indicators to measure the complete impact of articles from different perspectives in the web2.0 era. How to combine traditional indicators and Altmetrics indicators is a great problem to further study.

In the sample literatures, a large number of comprehensive Altmetrics indicators were constructed to reflect the social influence. However, these indicators are not as efficient as traditional indicators. The main reason is that these indicators are strongly dependent on using frequency of social software tools. Because some social tools are not popular for common users, the responding indicator value is very low and it is difficult to make a substantial contribution to the evaluation. In our researches, some tools, such as book reviewers, policy sources, peer review, Q&A thread and research highlight platform, provides only a small amount of data. How to choose reasonable Altmetrics indicator will be an important issue in the future studies.

It is important for us to note that Altmetrics has some limitations to evaluate social impaction. Firstly, it can only count Altmetrics data produced online, and cannot collect the data offline, such as the data in some printed media. Secondly, many Altmetrics tools are created around 2010, and the accumulative data is limited relatively to the traditional citation data. Finally, due to its geographical limitations, Altmetrics tools selected are used mainly in Europe and the United States, and less in non-English-speaking countries, so the collecting data are mainly concentrated in the Anglo-American countries. How to introduce the social tools of non-English-speaking countries to enrich the Altmetrics index is another problem to be solved.

On the one hand, new indicators and tools are needed to create to meet the diversification and the digitalization of scientific making and publishing. On the other
hand, it is possible for us to get online scientific communication and their evaluation data quickly and gives us more opportunities to create new evaluation indicators to improve the evaluation system of science and technology. Altmetrics is a very good example and its application has become an international trend to evaluate the social impact for academic literatures. It is also very likely that it will become an important criterion to assess the quality of scientific journals, papers and institutions. How to perfect the classification of Altmetrics indicators and how to construct a reasonable, effective and comprehensive evaluation system of literature impact still need more detailed and in-depth studies in large sample context in the future.

References
LI Gen. & MIAO Qi. (2016). Feature analysis and enlightenment of top 100 academic papers attracting the most attention online in 2014 listed on Altmetric. *ACTA EDITOLOGICA*, 2, 196-198.
TYPE: Posters

Analyzing the efficiency of research grants: a case study of grants issued by the National Natural Science Foundation of China

Hui Fang

fanghui@nju.edu.cn

State Key Laboratory of Analytical Chemistry for Life Science, School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, (China)

Introduction

One bibliometric topic in relation to grants is whether grant decisions are correct and fair. Nepotism has been found in the grant peer review system (Wennergås & Wold, 1997). Bornmann et al. (2008) found overestimation of approvals and underestimation of rejections in grant decisions. It is disputed to use applicant performance after receipt of grants to assess the correctness of grant decisions, because access to more abundant resources can explain why successful grant applicants on average produce more output than unsuccessful applicants (Melin & Danell, 2006). It is reasonable to compare researcher performance under identical funding conditions. Comparisons of PI performance should consider differences in grant size in financial terms. Measuring grant performance relative to grant size can help to identify researchers who use funds efficiently. It allows us to further investigate the distribution of fund utilization efficiency, and help to efficiently promote research development.

The grant efficiency of institutions and areas has been analysed by dividing their academic output by total financial input for a given year (e.g., Basu, 2014). However, this method ignores the delay in the transformation of inputs into outputs. In contrast, analysing funding efficiency based on inputs and outputs of individual grants can avoid errors arising from associating research outputs with inputs that did not support them.

Evaluating individual grants’ efficiency based on the articles they supported should also consider that many articles are supported by multiple grants. Each grant acknowledged in such an article only contributed partial financial support to it. In fact, the priorities of different funding bodies may overlap. Further, some authors took more money than they need, a practice known as double-dipping (Rigby & Julian, 2014).

This study evaluates grant efficiency arising from National Natural Science Foundation of China (NSFC) and exams grants’ efficiency changing with grant size. The NSFC supports basic research. We take the fact that articles may be supported by multiple grants into account, and avoid the delay in the transformation of inputs into outputs.

Data

On its website, the NSFC publishes detailed information on its grants. We chose to use medical sciences grants from the NSFC in this case study, as medical sciences are currently active research areas. Articles supported by the grants investigated can be retrieved from the WoS, which provides funding information on indexed articles published after August 2008.

This investigation was limited to grants starting in 2009, because articles supported by these grants were published after August 2008, and thus can be retrieved in the WoS. Many of these articles were published in the period 2009–2012. Grants that supported articles published after 2012 were removed, because we use a citation time window of at least 2 years in this preliminary study.

Methods

For articles supported by multiple grants, we assigned credit for an article equally to all grants supporting it. For an article supported by $g_{grant}$ grants and obtained $n_c$ citations, each grant is estimated as contributing $n_{ac}$ articles: $n_{ac} = 1/n_{grant}$, and obtaining $n_{cc}$ citations: $n_{cc} = n_c/n_{grant}$. $n_{ac}$ and $n_{cc}$ are defined as article credit and citation credit, respectively, that a given grant obtains from a specific article.

Each grant may support multiple articles. The total article credit of a given grant that support $n_a$ articles can be estimated as:

$$n_{ac} = \frac{1}{\sum_{i=1}^{g_{grant}} n_{grant}(i)}.$$  (1)

Its total citation credit can be evaluated as:

$$n_{cc} = \frac{n_c(i)}{n_{grant}(i)}.$$  (2)

PIs who have received larger grants have a higher probability of realizing larger achievements. So we define the article efficiency of a grant as its total article credit divided by its size $s_c$: $e_{ac} = n_{ac}/s_c$. Its impact efficiency is defined as its total citation credit divided by its size: $e_{cc} = n_{cc}/s_c$.

When evaluating the efficiency of a group of grants, we can integrate each grant’s outputs and size. The article efficiency of a group of $g$ grants can be estimated as:
\[ e_{\text{gps}} = \frac{\sum_{i=1}^{n_p} n_{\text{ts}}(i)}{\sum_{i=1}^{n_p} s(i)} \]  
\[ e'_{\text{gps}} = \sum_{i=1}^{n_p} n_{\text{ts}}(i) / \sum_{i=1}^{n_p} s(i) \]  

Its impact efficiency can be evaluated as:

\[ e'_{\text{gps}} = \sum_{i=1}^{n_p} n_{\text{ts}}(i) / \sum_{i=1}^{n_p} s(i) \]  

Results

We investigated 1,461 NSFC medical sciences grants issued starting in 2009, all of which supported articles published before 2013. This study included 8,988 articles. Individual grant sizes of the grants we investigated ranged from 80,000 to 10,000,000 RMB.

Figure 1 shows that a grant’s average total article credit and total citation credit increases with grant size. Certain grant sizes were uncommon, and associated with only a small number of grants. Grants of these sizes were combined with the closet grant sizes in one category using average size of the grants that were included to ensure that every point in Figure 1 represents at least 40 grants, except for the rightmost point, which represents only 34 grants.

![Figure 1. Change in total article credit ($n_{\text{tw}}$) and total citation credit ($n_{\text{tc}}$) with size of NSFC grant in medical sciences starting in 2009.](image)

Figure 2 shows the change in grant efficiency with grant size. In medical sciences, grants of 100,000 RMB or less have the highest efficiency in terms of both quantity and impact, and grant efficiency decreases with grant size before bottoming out, then rising, and finally stabilizing. The bottoming out occurs at 330,000 RMB. Each point in Figure 2 represents the same grants as the corresponding point in Figure 1.

![Figure 2. Change in grant article efficiency ($e_{\text{gps}}$) and grant impact efficiency ($e'_{\text{gps}}$) with size of NSFC grant in medical sciences starting in 2009.](image)

Discussion

Nowadays, there is an increasing concentration of resources among a smaller number of researchers (Bloch & Sørensen, 2014). It is thought that larger-sized grants, which foster the concentration of research resources partly at the expense of smaller grants, can improve overall scientific performance by funding top researchers. However, analysis shows that the National Science Foundation’s Small Grants for Exploratory Research program was highly successful in supporting research projects that produced transformative results (Wagner & Alexander, 2013). Our investigation shows that on average, small grants have the highest efficiency. It is consistent with the findings by Fortin and Currie (2013).

Figure 2 shows that large grants have higher efficiency than the medium-sized ones. One potential reason is that most PIs received large grants have high academic positions and reputations, which may help their articles be published more readily and obtain more citations.

References


The evolution of business model research (1993-2016): based on a co-word and maximum spanning tree analysis

Linqing Liu 1   Chang Tan 2   Wen Chen 3   Yiyang Lu 4

1 liulinqing@whu.edu.cn
Wuhan University ,Wuhan (China)

2 tanchangcw@whu.edu.cn
Wuhan University ,Wuhan (China)

3 Liaoning University, Shenyang (China)

4 Vanderbilt University, Tennessee (America)

Abstract
The aim of this study is to map and analyse the dynamic structure and the evolution of topics in business model research during the period of 1993-2016. Co-word analysis and maximum spanning tree analysis were employed to identify the main topics, while clustering analysis and strategic diagrams were applied to identify the research trends. The data set included 4134 articles that were grouped into three different periods: 1993-2005, 2006-2013, and 2014-2016, in order to reveal both how the results changed over time and what were the most recent research trends. The results showed business model literature had a stable diversity over time, and it was closely related to topics of electronic commerce in the early times, and it was closely related to topics of electronic commerce in the early times, and it was closely related to topics of electronic commerce in the early times, and it was closely related to topics of electronic commerce in the early times, and it was closely related to topics of electronic commerce in the early times, and it was closely related to topics of electronic commerce in the early times, and it was closely related to topics of electronic commerce in the early times, and it was closely related to topics of electronic commerce in the early times, and it was closely related to technology development played as an important driving force especially in recent literature.

Conference Topic
Co-occurrence analysis   Mapping and visualization   Co-word analysis

Introduction
The business model has attracted increasing attention from both academics and practitioners since the 1990s. As a new integrated interpretation of business value creation accompanying the thriving of electronic commerce, business model construct has occupied the core logic of the firms (Linder & Cantrell, 2000). And it continually expands it framework to account for corporate growth and success, as well as influence the whole business picture. The academic field has examined the business model through different lenses and varying aspects. These areas include business economics, computer science, telecommunications, and operation management and so on. Part of this diversity can be attributed to the cognitive difference in the very concept, on which consensus is hard to reach in terms of its definition, nature, structure, and evolution (Morris, Schindehutte, & Allen, 2005). Therefore, a co-word analysis can provide an important unifying effort to identify the main characteristics and focal points of business model research as well as the relationship and structure among business model research topics. It will also help to grasp the directions and trends of this research area in its continual development.

Related Studies

Business model concept.
The business model had long been integral to economic behaviour, but its concept didn’t become prevalent until the increasing popularity of the Internet in the mid-1990s, and it has
been gaining pace since then (Zott, Amit, & Massa, 2011). The earlier definition of business model was “an architecture for the product, service and information flows, including a description of the various business actors and their roles, of the potential benefits for the various business actors, and of the sources of revenues” (Timmers, 1998). Later Amit & Zott(2001) placed more emphasis on the value aspect and referred to the business model as “the design of transaction content, structure, and governance so as to create value through the exploitation of business opportunities”. Morris et al.(2005) went on to identify six fundamental components of business models: value proposition, customer, internal processes/competencies, external positioning, economic model, and personal/investor factors. A more recent definition depicted the business model as “a reflection of the firms realized strategy (Johnson, Christensen, & Kagermann, 2008)”.

These definitions have some common key points, such as the stress of value, including value creation and value capture, and the stress of systematic nature of business model components as a whole. However, there isn’t one generally accepted definition of ‘business model’ (Morris et al., 2005), and this lack of unanimous definitions is somewhat a potential source of confusion, dispersing rather than converging the researches on business models (Zott et al., 2011).

**Business model related research area**

On the whole, the current business model research spreads around two areas: business model ontology and business model design. Business model ontology studies seek to uncover the intrinsic quality of business models and reach a unified logic for them, mostly through case study or literature review. Some simply described the essence of a business model in the logic of value creation, capture, delivery and conversion (Teece, 2010), while some identified several orders of themes and structure the concept into levels of decomposition with increasing complexity and depth (Osterwalder, 2004). Business model design researches seek to achieve growth and generate revenues for a specific type of company or market. These researches cover a wide range of industries including electronic commerce, airline carriers, hospitals, and pharms, Internet of things, banks, scholarly publishing and energy systems and so on, and they deal with nearly all segments inside a company or market, varying from production to logistics, from pricing strategy to financial sustainability, and from technology innovation to corporate social responsibility. Timmers(1998) identified eleven business models of electronic commerce and examined their degree of innovation, among which e-shop was the most innovative while value-chain integrators was the least innovative. In order to replace the prevailing profit-based journal publishing model, Fyffe and Shulenburger(2002) introduced a BioOne business model to meet the interlocking goals of both providing broad and enduring access and supporting the publishing enterprise of scholarly societies. And Tan and Yavuz(2015) presented a new business model design to offer energy saving technologies as a service, it captured energy efficiencies improvements, energy consumption variation, energy prices uncertainty and carbon offsets revenues, thus leading to a win-win situation for business and the environment.

Despite the twenty years’ development and many excellent research papers in business model literature, we found that most studies dealt with a very specific and rather limited research area, and few focused on the whole structure and topic correlations of business model research, nor did they systematically examine the evolution of this literature. This study aims to use co-word analysis to uncover the hidden information of various business model research areas since its existence, including relationships of keywords, research structure and research advances.

**Co-word analysis and maximum spanning tree analysis**

Co-word analysis is an important bibliometric way to reveal the relationships of concepts and ideas within a specific scientific domain. Co-word analysis rests on the assumption that
extracted high-frequency keywords could represent a specific research subject and direction of a field (Callon, Courtial, & Laville, 1991). It also calculates the co-occurrence frequency of each word pair (Yan, Lee, & Lee, 2015b). Higher co-word frequency means stronger correlation in keyword pairs, and a cluster of keywords based on similarity of these keyword pairs can be understood as a short description of a research theme (Hu, Hu, Deng, & Liu, 2013; Huang et al., 2015).

Maximum spanning tree (MST) is an algorithm that can be applied to co-word analysis to generate a network representation of the similarity matrix, which helps to develop intuition about the structure of the specific scientific domain as well as the dynamics of keywords. It was used by Hidalgo, Klinger, Barabási, and Hausmann (2007) to identify industry dynamics in product space and reveal the transformation path of products. This method gives us a quick visual way to show the most relevant links and to determine where the keywords are located and where they could be headed in the network. It also gives us an idea of core-periphery structure of the given network. The basic principal of MST approach is to generate a “skeleton” of the whole network, which contains a sum of weights that is maximal.

While co-word analysis was rarely used in business model domain, we then applied this method together with maximum spanning tree analysis to map the evolution of structure and hot topics and to examine the development tendencies in this research area.

**Research methods**

**Data Collection**

In this study, we chose Science Citation Index Expanded (SCI-Expanded), Social Sciences Citation Index (SSCI) and Arts & Humanities Citation Index (A&HCI) of the institute for Scientific Information (ISI) Web of Science as the data source. The data were collected in January 3rd 2017. Data retrieval strategy: the theme is “business model” or “business models”; the publication period is from 1993 to 2016. To make data more accurate, we confined the reference type to “articles” and language “English”. Finally, a total of 4134 records were obtained.

Figure 1 shows the yearly output of business model research. The output distribution has periodical characteristics: there were only 53 records before 1998, and the year of 1999 saw a start of obvious growth, which reaches its peak in 2005. This growth period resulted in 736 records. After a small decrease in 2006, there was an accelerated growth from 2006 to 2013, with a total records of 1991. The records of 2014 were almost the same compared to 2013, and the recent three years have a total records of 1354.

![Figure 1. The output of business model research during 1993-2016](image)
In order to study the evolution of the topics and the recent research trends in the literature, we divided the records into three sub-periods based on the output periodical growth characteristics and the number of target documents published per period. The three sub-periods are 1993-2005, 2006-2013, and 2014-2016. Following Dehdarirad, Villarroya and Barrios(2014), we fixed a longest first sub-period in order to get a representative number of published papers and keywords, and a relatively short last period to reveal recent research fronts. The first period (1993-2005) spans 13 years (and includes 789 records, accounting for about 19% of the total records), the second (2006-2013) and third periods (2014-2016) spans respectively 8 years (1991 records: 48%) and 3 years (1454 records: 33%). Figure 2 shows the output of business model research in each period.

![Figure 2. The output of business model research in three periods](image)

**Data process**

The co-word analysis conducted in this study involved six sequential steps, as shown in Figure 3. They are data extraction, keywords selection, matrices construction, MST analysis, clustering and visualization. Firstly, for each period, keywords were extracted from papers in the specific time span using bibexcel software, and we got 1685 keywords from 789 papers in period 1, with an average 2.13 keywords for each paper; we got 3286 keywords from 1991 papers in period 2, with an average 1.65 keyword for each paper; we got 2754 keywords from 1354 papers in period 3, with an average 2.03 keywords for each paper.

Secondly, for each period, we eliminated invalid keywords and standardized some keywords by combining synonyms to achieve more precise results, E.g. “e-business” and “e-commerce” were replaced by “electronic commerce”; “low cost carrier” and “low cost airlines” were replaced by “low-cost airlines”; “base of the pyramid”, “BOP” and “Bottom-of-the-pyramid” were replaced by “bottom of the pyramid”. Then, g-index method was used to select the high frequency keywords: If it ranks in decreasing order of the number of keywords frequency, the g-index is the (unique) largest number such that the top g keywords receive (together) at least g2 frequency(Yan, Lee, & Lee, 2015a). In our case, the g-index was respectively 14, 28 and 22 for period 1, 2 and 3.

Thirdly, for each period, we obtained a co-occurrence matrix of these high frequency keywords with Bibexcel software. On the basis of the three co-occurrence matrices, we calculated the
similarity index between two words according to the study of Hidalgo et al. (2007) with Excel. The formula is as below:

$$\Phi_{ij} = \min\{P(nR_{ij}|nR_i), P(nR_{ij}|nR_j)\}$$

$\Phi_{ij}$ is the similarities between keyword i and j; $nR_i$ is the frequency of keyword i; $nR_j$ is the frequency of keyword j; $nR_{ij}$ is the co-occurrence frequency of keyword i and j.

---

**Figure 3. Data process flow**

Fourthly, for each period, we performed MST analysis on each similarity matrix in order to get the “skeleton” of the network that shows where the keywords are located and where they could be headed in the network. Instead of using binary values of the matrices, MST analysis takes every similarity value into consideration, thus offering one new and useful angle to look at the pairwise relationships and whole network relationships of these keywords. For the first step, we considered the strongest non-diagonal value of the similarity matrix and then selected the strongest link connected to that dyad. We then picked up the strongest link connecting a new node to our triad and continued adding links until all the nodes on the network were included (Hidalgo et al., 2007). The above process was completed using Java, and Netdraw software was then applied to show the three MST networks.

Fifthly, for each period, we conducted clustering analysis to classify keywords into different groups. Clustering analysis is an exploratory tool to group relatively homogenous observations or cases. Each cluster thus describes the class to which its members belong with the application
of different techniques (Yan et al., 2015b). There are a variety of methods for clustering analysis, such as hierarchical clustering, Girvan-Newman clustering (Girvan & Newman, 2002), density-based clustering (Sander, Ester, Kriegel, & Xu, 1998) and so on. We used combinatorial optimization clustering in Ucinet software to classify the keywords in each of the three period. Combinatorial optimization is about finding an optimal object from a finite set of objects. It deals with problems where exhaustive search is not feasible and the set of feasible solutions is discrete (Du & Pardalos, 2013).

Finally, for each period, strategic diagrams were drawn in order to identify and visualize the status and evolution trends of clusters in business model research. A strategic diagram is a two-dimensional space with abscissa axis represents the centrality and the ordinate axis represents the density (Dehdarirad et al., 2014). Centrality is used to measure the correlation degree among different clusters, while density is to measure cohesiveness within one specific cluster (Hu et al., 2013). In strategic diagram, the origin is the median of these two axis, and there are four different quadrants accordingly. Research topics in quadrant I are considered to be well-developed with relatively high centrality and density, while in quadrant II, they only have well-developed internal ties but unimportant external ties, which indicates they are specialized and peripheral in nature. In quadrant III, research topics have both weak centrality and weak density, these topics are considered as either emerging or disappearing, and in quadrant IV, and research topics have high centrality but low density, indicating they are general or basic topics in literature (Dehdarirad et al., 2014). In our study, the density and centrality were first calculated using Ucinet software, and the strategic diagram was then displayed using Excel.

Results and discussion

Research direction distribution analysis

In Table 1, the top 10 research directions in business model area during each of the three period are presented. For all the three periods, Business & Economics, Engineering and Computer Science always occupies the first three places in the rankings, while in period 1 Computer Science instead of Business & Economics was the most popular research direction. This indicated business model research is the most productive in these three directions and greatly related with them as well. However, Health Care Sciences & Services and Communication in period 1 fell out of top 10 during period 2 and 3, while Environmental Sciences & Ecology and Energy & Fuels were emerging and gaining pace in the last two periods. The rest six research directions remained in the top 10 rankings throughout the three periods, which suggested that business model research spread across broad yet stable research directions in literature during these years.

<table>
<thead>
<tr>
<th>No.</th>
<th>Publication</th>
<th>Research Direction</th>
<th>Publication</th>
<th>Research Direction</th>
<th>Publication</th>
<th>Research Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>296</td>
<td>Computer Science</td>
<td>760</td>
<td>Business &amp; Economics</td>
<td>529</td>
<td>Business &amp; Economics</td>
</tr>
<tr>
<td>2</td>
<td>220</td>
<td>Business &amp; Economics</td>
<td>382</td>
<td>Engineering</td>
<td>243</td>
<td>Engineering</td>
</tr>
<tr>
<td>3</td>
<td>202</td>
<td>Engineering</td>
<td>345</td>
<td>Computer Science</td>
<td>170</td>
<td>Computer Science</td>
</tr>
<tr>
<td>4</td>
<td>119</td>
<td>Telecommunications</td>
<td>152</td>
<td>Information Science &amp; Library Science</td>
<td>165</td>
<td>Environmental Sciences &amp; Ecology</td>
</tr>
</tbody>
</table>
Frequency analysis

The results of the frequency analysis in the three periods are shown in Table 2. In period 1, “Electronic commerce” and “Internet”, the thriving of which invited increasing attention to business model research, were among the highest frequency keywords. “Electronic markets”, “Pricing” and “Security”, which were related to E-commerce studies (Yan et al., 2015a), were also high-frequency keywords in business model research. “Low-cost airlines” was another business model that received attention besides electronic commerce in period 1.

In period 2, “Electronic commerce” and “Internet” were still very hot topics in the literature. However, there saw the emerging of three new keywords that were never seen in business model research in period 1: “Cloud computing” and “Web 2.0” reflected the development and application of technology, and “Open innovation” was a new way of thinking in business model innovations. Moreover, keywords such as “Sustainability”, “Strategy”, “Entrepreneurship”, “Intellectual property” and “Regulation” received much more attention in period 2 compared to period 1.

In period 3, the most recent trends showed “Electronic commerce”, with a rank of 19 in frequency, was not that popular anymore, and “Internet” even fell out of the high-frequency keywords. Their places were replaced by “Sustainability” and “Business model innovation”. New emerging keywords in period 3 were “Big data” and “Internet of things”, reflecting the further development of technology. Besides, two of the three emerging keywords in period 2 still maintained high frequency in period 3, namely “Cloud computing” and “Open innovation”. Keywords like “Servitization” “China”, “Electric vehicles” and “Renewable energy” received much more attention in period 3 compared to the other two periods.

Table 2. High frequency keywords in the three periods during 1993-2016

<table>
<thead>
<tr>
<th>No.</th>
<th>Frequency</th>
<th>Keywords</th>
<th>Frequency</th>
<th>Keywords</th>
<th>Frequency</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>52</td>
<td>Business model</td>
<td>291</td>
<td>Business model</td>
<td>95</td>
<td>Business model</td>
</tr>
<tr>
<td>2</td>
<td>47</td>
<td>Electronic commerce</td>
<td>62</td>
<td>Innovation</td>
<td>47</td>
<td>Sustainability</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>Internet</td>
<td>41</td>
<td>Electronic commerce</td>
<td>40</td>
<td>Business model innovation</td>
</tr>
</tbody>
</table>

1727
<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>Innovation</td>
<td>31</td>
<td>Internet sustainability</td>
<td>34</td>
<td>Innovation entrepreneurship</td>
<td>Cloud computing servitization</td>
</tr>
<tr>
<td>11</td>
<td>Supply chain management</td>
<td>29</td>
<td>Strategy</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Knowledge management</td>
<td>27</td>
<td>Case Study</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Electronic markets</td>
<td>23</td>
<td>Cloud computing</td>
<td>17</td>
<td>China</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Case study</td>
<td>22</td>
<td>Business model innovation</td>
<td>17</td>
<td>Case study</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Pricing</td>
<td>22</td>
<td>Supply chain management</td>
<td>16</td>
<td>Value creation</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Information technology</td>
<td>17</td>
<td>Open innovation</td>
<td>16</td>
<td>Open innovation</td>
<td>Electric vehicles renewable energy strategy</td>
</tr>
<tr>
<td>6</td>
<td>Value chain</td>
<td>17</td>
<td>Entrepreneurship</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Low-cost airlines</td>
<td>17</td>
<td>Intellectual property</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Security</td>
<td>16</td>
<td>Regulation</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Value creation</td>
<td>15</td>
<td>Web 2.0</td>
<td>14</td>
<td>Big data corporate social responsibility</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>Supply chain</td>
<td>14</td>
<td>Internet of things</td>
<td>Bottom of the pyramid</td>
<td>Electronic commerce sustainable development collaboration</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>Networks</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>Globalization</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>Collaboration</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>Business strategy</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>Convergence</td>
<td>10</td>
<td>Energy efficiency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>China social networks</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Corporate social responsibility</td>
<td>11</td>
<td>Information systems</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Services</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Value chain</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Throughout the three periods, “Electronic commerce”, “Innovation”, “Case study” and “Value creation” were always high-frequency keywords. The other three keywords had a relative stable rank in frequency except “Electronic commerce”, which as noted before experienced dramatic frequency decline in period 3. It indicated that business model research had a broad and diverse topics and it has close relation with technology development.

Correlation matrix
Based on the co-occurrence matrix calculated through Bibexcell, we used conditional probability method according to Hidalgo et al. (2007) to measure similarity between two keywords. Two keywords are more related if the pairwise similarity index is higher. Similarity matrix is not presented here for limited space.

MST analysis
After obtaining the similarity matrices, we performed MST algorithm using Java on each matrix and visualized the results using Netdraw software, as shown in Figure 4, 5 and 6. The nodes represented keywords and the links represented similarity between these keywords. The width of links within each figure was determined by the relative value of similarity between the two nodes, and two keywords that are more related have a wider line. MST analysis, which contains a sum of similarities that is maximal, offered one new and useful angle to look at the pairwise relationships and whole network relationships of these keywords in three periods. Firstly, “Business model” varied in network positions within the literature throughout the three periods. In period 1, it was related closely with “Electronic commerce”, which was one important driving element to speed up business model research during 1993-2005; but in period 2, “Business model” was more closely related to “Innovation” in the MST network, and this relationship continued in period 3, while “Business model” was closely connected with “Sustainability”, “Value creation”, and “Entrepreneurship” as well. It indicated more direct aspects of business model had been studied and rooted in the literature since its emergence with ecommerce.

Secondly, for the newly emerging keywords in period 2 and 3, their paths to enter the literature could be revealed. In Figure 2, “Cloud computing” and “Web 2.0” were respectively closely connected with “Services” and “Collaboration”, while “Open innovation” was highly related to “Intellectual property”; In period 3, “Big data” and “Internet of things” appeared to connect closely with each other, while “Big data” was also deeply related to “Cloud computing”.

Thirdly, in the MST networks, keywords with more links with other keywords were more popular or more deeply rooted in the literature during each period (Hidalgo et al., 2007). In our analysis, one salient keyword in the literature needed to satisfy two conditions: 1. it had more than 2 links with other keywords; 2. it ranked above (or equaled) middle-place in frequency of keywords of each period. So in period 1 the salient keywords were “Electronic commerce”, “Electronic markets” and “Supply chain management”, and in period 2, the salient keywords were “Internet”, “Strategy” and “Innovation”. Salient keywords in the most recent period appeared to be more diverse, they were “Business model”, “value creation”, “open innovation”, “Entrepreneurship” and “China”.

Finally, a keyword located in the periphery of the MST network with the weakest link was loosely connected to the literature, which indicated its less importance in the existing literature compared to others. “Low-cost airline” in period 1, “Supply chain management” in period 2 and “Bottom of pyramid” in period 3 were such keywords. As “Supply chain management” was a salient keywords in period 1, it experienced a dramatic popularity decline in period 2.
Figure 4. MST of keywords in period 1(1993-2005)

Figure 5. MST of keywords in period 2(2006-2013)

Figure 6. MST of keywords in period 3(2014-2016)
Combinatorial optimization clustering was performed to group the keywords in each of the three periods into different groups with Ucinet software, and the results are shown in table 3-5. In order to achieve deeper understanding of each cluster, we calculated the total word frequency and average frequency as well (also in table 3-5). Keywords in each cluster can reflect the corresponding research directions, and keywords with more frequencies attracted more attentions (Hu et al., 2013).

In period 1, the keywords were divided into three clusters with a fit value of 0.399 and r-square value of 0.361. In table 3, Cluster 1-1 was highly related to electronic commerce, Cluster 1-2 was more about supply chain management, and Cluster 1-3, which included research topics as “Electronic markets”, had some overlaps with Cluster 1-1, but discussed more about management issues such as “Knowledge management”, “Pricing” and “Security” as well. Overall, Cluster 1-1 had a much higher average frequency than Cluster 1-2 and Cluster 1-3, so in period 1, business model research had devoted its greatest attention to electronic commerce (Zott et al., 2011).

In period 2, the keywords were divided into six clusters with a fit value of 0.373 and r-square value of 0.394. In table 4, Cluster 2-1, with an average frequency of 73.33, was the most popular cluster in period 2, and it included two salient topics discussed in MST analysis: “Innovation” and “Strategy”. Cluster 2-2 and Cluster 2-5, with relatively lower average frequencies compared to the other four, were respectively represented by “Web 2.0” and “Cloud computing”, which as discussed before were two new developments in the literature driven by technology advances in period 2. Cluster 2-3 was related to internet and business model innovation, and Cluster 2-4 was about electronic commerce. Cluster 2-6 focused on strategy as it includes topics as “Case study”, “Business strategy” and “Value chain”. So for business model research in period 2, while it still emphasized on electronic commerce, strategy and innovation related topics received increasing attention and web 2.0 and cloud computing related topics were newly emerging.

In period 3, the keywords were divided into five clusters with a fit value of 0.431 and r-square value of 0.324. In table 5, Cluster 3-1 attracted the most attention from recent literature with an average frequency of 36.33, and its representative keywords were “Sustainability” and “Innovation”. Cluster 3-2 had the lowest average frequency of 15.00, but it included the newly emerging topics in period 3: “Big data” and “Internet of things”; this cluster indicated the technology development trends in recent business model literature. Cluster 3-3 and Cluster 3-4 were respectively about entrepreneurship and business model innovation, and Cluster 3-5 included some other topics besides the above such as the emerging market of China and the declining topic of electronic commerce. So for the most trends in business model research of period 3, while innovation and technology development was still an important part, entrepreneurship related topics attracted more attention and topics tend to becoming more diverse.

In each of the three periods of business model research, there were some overlaps between clusters, which increased the difficulty to define its subfields in this already diverse literature. However it was also this interlink that made this literature integrated and drive the literature move forward.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Keywords</th>
<th>Number of keywords</th>
<th>Total frequency</th>
<th>Average frequency</th>
<th>Centrality</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Business model(52); Electronic commerce(47); Internet(17);</td>
<td>5</td>
<td>136</td>
<td>27.20</td>
<td>13.035</td>
<td>0.061</td>
</tr>
<tr>
<td>Cluster</td>
<td>Keywords</td>
<td>Number of keywords</td>
<td>Total frequency</td>
<td>Average frequency</td>
<td>Centrality</td>
<td>Density</td>
</tr>
<tr>
<td>---------</td>
<td>---------------------------------------------------------------------------</td>
<td>--------------------</td>
<td>-----------------</td>
<td>-------------------</td>
<td>------------</td>
<td>---------</td>
</tr>
<tr>
<td>1-2</td>
<td>Innovation(14); Low-cost airlines(6); Supply chain management(11); Case study(7); Information technology(6); Value chain(6); Value creation(5)</td>
<td>5</td>
<td>35</td>
<td>7.00</td>
<td>12.049</td>
<td>0.054</td>
</tr>
<tr>
<td>1-3</td>
<td>Knowledge management(9); Electronic markets(7); Pricing(6); Security(5)</td>
<td>4</td>
<td>27</td>
<td>6.75</td>
<td>12.046</td>
<td>0.065</td>
</tr>
</tbody>
</table>

Table 4. Six clusters of business model research in period 2

Table 5. Five clusters of business model research in period 3
<table>
<thead>
<tr>
<th>Cluster</th>
<th>Keywords</th>
<th>Number of keywords</th>
<th>Total frequency</th>
<th>Average frequency</th>
<th>Centrality</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-1</td>
<td>Business model(95); Sustainability(47); Innovation(34); Value creation(16); Renewable energy(14); Sustainable development(12)</td>
<td>6</td>
<td>218</td>
<td>36.33</td>
<td>8.412</td>
<td>0.049</td>
</tr>
<tr>
<td>3-2</td>
<td>Cloud computing(20); Open innovation(16); Big data(14); Internet of things(14); Collaboration(11)</td>
<td>5</td>
<td>75</td>
<td>15.00</td>
<td>7.463</td>
<td>0.072</td>
</tr>
<tr>
<td>3-3</td>
<td>Entrepreneurship(24); Corporate social responsibility(14); Bottom of the pyramid(13); Energy efficiency(10)</td>
<td>4</td>
<td>61</td>
<td>15.25</td>
<td>4.425</td>
<td>0.028</td>
</tr>
<tr>
<td>3-4</td>
<td>Business model innovation(40); Servitization(18); Case study(17); Strategy(10)</td>
<td>4</td>
<td>85</td>
<td>21.25</td>
<td>6.772</td>
<td>0.048</td>
</tr>
<tr>
<td>3-5</td>
<td>China(17); Electric vehicles(16); Electronic commerce(13)</td>
<td>3</td>
<td>46</td>
<td>15.33</td>
<td>4.462</td>
<td>0.039</td>
</tr>
</tbody>
</table>

**Visualization**

The centrality and density calculated for each cluster in three periods were shown in table 4, 5 and 6, and the strategic diagrams on this basis were shown respectively in figure 7, 8 and 9. In each diagram, the clusters were represented by spheres of different size, which were proportional to the number of keywords in each cluster. From the strategic diagrams, the position and evolution of clusters, as well as the trends of current research can be revealed.

In period 1, as shown in figure 7, quadrant I includes cluster 1-1, indicating that electronic commerce related topics were well-developed in period 1. And it could be considered the core trend of this period. Cluster 1-3 fell into quadrant with relative high density but low centrality, indicating that these topics were rather specialized and located peripherally in period 1. Cluster 1-2 in quadrant III had both low centrality and density, and it was regarded as either emerging or disappearing topics in period 1.

In period 2, as shown in figure 8, quadrant I includes cluster 2-1, cluster 2-2 and cluster 2-3. Cluster 2-1, which includes two salient keywords of “Innovation” and “Strategy” and have highest average frequency as discussed before, was very well-developed, and could be viewed as the core trend of this period as its centrality ranked top to indicate its wide connection to other clusters. Cluster 2-2, although had low average frequency, had a relative high density and centrality, indicating its deep root in the literature with its emphasis on technology and networks. Cluster 3-3 with the salient keyword “Internet” in period 2, also had close internal connections and wide external connections in the literature. In quadrant II, Cluster 2-5 as represented by “Cloud computing” seemed rather specialized and was still in the periphery of the literature in
this period. Cluster 2-4 and Cluster 2-6 were in quadrant III, as discussed before, electronic commerce related topics were slowly losing its popularity since period 2, while other topics were still emerging.

In period 3, as shown in figure 9, Cluster 3-1 with the highest centrality and Cluster 3-2 with the highest density were both in quadrant I, indicating they were well-developed. As Cluster 3-1, represented by “Sustainability” and “Innovation”, had the most external connections and Cluster 3-2, representative of recent technology developments, had the most internal connections, they could be core trends of this period together. Cluster 3-4 was in quadrant I as well, but close to quadrant IV, and Cluster 3-3 and Cluster3-5 were in quadrant III, indicating they were either emerging or disappearing in period 3.

Figure 7. Strategic diagram for period 1(1993-2005)

Figure 8. Strategic diagram for period 2(2006-2013)

Figure 9. Strategic diagram for period 3(2014-2016)
Conclusion

In this study, we conducted co-word analysis together with maximum spanning tree analysis (Hidalgo et al., 2007) to describe the evolution and current state of the business model research by dividing the literature into three sub-periods (i.e., 1993-2005, 2006-2013, 2014-2016). Through using the tools of bibexcel, Java and Ucinet, we obtain some clear analysis results of business model research.

1) In terms of evolution of the number of documents, the results revealed that about one third of the total body of literature was published in the recent three years, suggesting the current rich interest of researchers on business model research. Moreover, business model research spread across broad yet stable research directions throughout the three periods, indicating the stable diversity of this area.

2) Frequency analysis reveals the evolution of hot topics in the literature. Throughout the three periods, “Electronic commerce”, “Innovation”, “Case study” and “Value creation” always maintained hot-topic status, although “Electronic commerce” started to experienced dramatic frequency decline since period 2. In terms of research methods, case study was obviously important and frequently used in this literature. Period 2 had new emerging topics of “Open innovation”, “Cloud computing” and “Web 2.0”, while period 3 had “Big data” and “Internet of things”.

3) MST analysis reveals the evolution of pairwise relationships between high frequency keywords. As “business model” was connected most closely to keywords from “Electronic commerce” to “Innovation”, and finally to a keyword group of “Sustainability”, “Entrepreneurship” and so on, the literature has had more direct aspects of business model studied and rooted over time since its emergence accompanying electronic commerce. Period 1 had salient topics of “Electronic commerce”, “Electronic markets” and “Supply chain management”, and period 2 had “Internet”, “Strategy” and “Innovation”, while salient topics more recently appeared to be more diverse, they were “Business model”, “value creation”, “open innovation”, “Entrepreneurship” and “China”. By contrast, “Low-cost airline” in period 1, “Supply chain management” in period 2 and “Bottom of pyramid” in period 3 were loosely connected to the body and in the periphery of literature in each period.

4) Cluster analysis and visualization analysis using strategic diagram reveals the evolution of topic clusters in business model research over time. Keywords in period 1 could be divided into 3 clusters and the core trend was cluster 1-1 represented by “Electronic commerce”; keywords in period 2 could be divided into 6 clusters and the core trend was cluster 2-1 represented by “Innovation”, “Sustainability” and “Strategy”; keywords in period 3 could be divided into 5 clusters and the core trends were cluster 3-1 represented by “Innovation” and “Sustainability” and cluster 3-2 representing technology development in business model research. For the main body of business model literature, it showed that business model first attracted research attention mostly due to the rise of electronic commerce, but continued its development into innovation, sustainability and strategy related areas, with the driving force of technology development such as the emerging of cloud computing techniques and web 2.0 network techniques, and most recently, the research trends no longer emphasized electronic commerce related topics, but had embraced technology development related topics as a major research area besides the already well-developed areas as “innovation” and “sustainability”.

In summary, this study reveals the research advances of business model research in different time periods, and helps us grasp the current status and trends by reasonable interpretation of the results. The analysis across three periods is necessary and important for understanding the dynamic structure of the literature, for it can provide information about specific transient trends (Dehdarirad et al., 2014), and can be used to identify gaps to promote research in emerging areas in the future.
Acknowledgments
The financial support from China Scholarship Council (file no.201606270047) is greatly acknowledged.

References
Research of the Role of Leading Enterprises in Promoting Industrial Technology Development based on Patent Content Analysis

Wang bo¹², Liu Shengbo², Ding Kun², Liu Zeyuan²

(1. School of Management, Dalian Polytechnic University, Dalian, 116034
2. WISELab, Dalian University of Technology, Dalian 116023)

Abstract: This paper takes “technical term” of patent literature as basic analysis unit. Based on an in-depth study of patent texts, a quantitative description of leading enterprises’ patented technology and industry technology standards as well as the entire industrial technology development has been made. Taking the patents and essential patents of communication industry in recent 24 years as samples, this paper reveals the function mechanism of industry leaders in the process of industrial technology development via a similarity comparison between leading enterprises’ patented technologies and industry technology standards and industrial patented technologies. Results indicate that technologies of leading enterprises are ahead of leading industrial technology of the same period and compose an important part of the contemporaneous pioneering technology thus providing important technological base for the formation of new technological paradigms. Meanwhile, leading enterprises in communication industry take an active part in standardization work and possess a large number of standard patents during the process of industrial standardization, which promotes significantly the formation of both industrial technology standards and technological paradigms.

Keywords: Leading enterprises; Technology standard; Industrial technology development; Patented technology; Content analysis

1 Introduction

With the arrival of the era of knowledge and information based economy, enterprise and industry technology analysis based on patentometrics also rose gradually (Zou and Li 2002). Griliches (Griliches 1998) even point out that patent is the unique resource of information which can timely identify the process of technical change. Nothing else even comes close in the quantity of available data, accessibility, and the potential industrial, organizational, and technological detail. Patent, as an important carrier to convey technical knowledge, has been introduced to the field of technology development by its objectiveness, availability and timeliness. In recent years, patent data has been widely used to measuring innovation econometrically in both firm and industry level.

Some scholars started to explore the development route of industrial technology by analyzing the development of enterprises’ patented technology within the industry. For example, Holger Ernest made an analysis of the technical development of NC machine tool industry based on patent data (1984-1992) of 50 German machine tool companies (Ernst 2001). Besides, after studying the development of patented technology of Fortune Global 500 Companies, Xianwen Wang (Wang, Zeyuan et al. 2010) made a research of the relationship of technological development among different industries with reference to their core competence in patented technology and competitive situation. Furthermore, Dehai Guo (Guo 2008) analyzed the patented technology...
distribution of multinational pharmaceutical companies in China. He made it clear that how foreign medical companies developed their patent in China and proposed some advice for the development of traditional Chinese medicine in industrial patented technology. At the same time, the analysis of the impact exerted by technical standard on industrial revolution and development is also on the rise, including its impact on the choice of leading industry planning and the interrelationship between technical standard and technological revolution of the industry (Bekkers, Iversen et al. 2006; Bekkers and Martinelli 2010).

Most of the study regarding the relationship between enterprise patented technology and industry technology development is based on structured patent information such as number of patents, patent classification and patent citation. Meanwhile, researchers acknowledge that unstructured information with patent content, like patent abstract, patent claims and patent specification contains more valuable information, which could depict the development of technology more directly, accurately and intuitively. Few analyses deeply into patent texts of unstructured information such as abstracts, independent claims and specification have been made.

Meanwhile, technology standardization has always been regarded as a powerful weapon in international market competition. With the emergence of anticipatory standards, the improvement of the standardization bodies’ database and the real-time update of standardization information, technology standardization shows an obvious forward-looking of the industrial technology development. As technical standard tends to be patented, patented technology standard gradually break the scope of traditional technical specification and plays an important role in international market competition and normalizing industrial technology order as well as promoting industrial technology development.

Since patent tends to be standardized, traditional analysis methods and tools can hardly reflect the relationship between standard and patent accurately and intuitively (Feng and Zhang 2011). In researches about leading enterprises’ patented technology and the relationship between technical standard and industrial technology development, what role do leading enterprises play in the process of industrial technology development? How? With traditional analysis methods and measurement of structured patent information only, it is rather difficult to make a clear description of the technological development trace of leading enterprises and the industry, not to mention to figure out the function mechanism of leading enterprises’ development on industrial technology development.

With the constant improvement of patent database electronization and the development of text mining technology, makes the in-depth analysis into the patent content level feasible. This paper takes “technical term” of patent literature as basic analysis unit and analyses deeply into patent texts by making use of Natural Language Processing, text mining technology and patentometrics so as to achieve a quantitative description and a comparative analysis of industrial technology and leading companies’ patented technology content. Theoretically, this study could offer new theoretical perspectives and analyzing thoughts for figuring out that how leading enterprises’ technological development promote industrial technological development so as to get a quantitative, deep and accurate understanding of the general development process of industrial technology. Practically, by analyzing how leading enterprises function in the process of industrial development through technical standard, this paper explores factors that determine industrial technology development and the rules beneath, thus being significant for leading enterprises playing their role and achieving innovation-driven development in China.
2 Data Sources and Methodology

2.1 Data Sources and Retrieval of Communication Industrial Patent

In this paper, all the patent data of communication industry was taken from Derwent Innovation Index (DII). DII has recorded more than 10 million basic patents and about 30 million patents, which come from over 40 patent institutes all over the world since 1963, and updates at a speed of over 25000 patents per week. Besides, Thomson Scientific organizes about 350 experts every year to make a further processing of patents in DII. They index each patent text of different countries, remove those obscure terminologies of the original patent texts and rewrite or extend those titles and abstracts in English with commonly used technical words. The adoption of this database not only increases the extraction efficiency of technical terms but also makes the extraction results of patent subject terms more meaningful, which will directly affect the mining results of technical subjects and technical intelligence hidden in patents.

As for the retrieval of communication patents, this paper adopted a combined retrieval strategy based on subject terms and classification codes. Taking core patent set of international communication technology standard as the test set, it was found that almost all core communicational technology focused on Derwent patent classification code such as W01, W02, T01, W04 and W05 and International Patent Classification Code such as H04 and that there was a strong cooccurrence relation in those technology fields. Figure 1 shows how the test was conducted in details:

![Diagram showing the retrieval process](image_url)
In Figure 1, Q represents training set consisting of all testing patents QA. Firstly, after making a statistical analysis of the International Patent Code (IPC) and Derwent Manual Code (DC) in test set, a frequency of occurrence ranking of IPC and DC is got and store the Top 1 IPC and MC. Then search patents that contain Top 1 IPC number in all testing patents and get a patent dataset Qtop which is removed later from the training set Q thus getting a new patent dataset QN. If QN makes up more than 5% of QA, the new dataset QN will be taken as training data set (Q=QN) and a new cycle begins from re-ranking the frequency of occurrence of IPC and DC in training set. In contrast, a proportion less than 5% for QN in QA means that the patents corresponding to selected classification codes account for over 95% of the total amount.

Through a retrieval test on combined classification codes, a retrieval strategy was finally defined as follows: TS=communication and (IP=H04* and DC= (W01 or W02)) with a time span from 1990 to 2014 and index database including CDerwent, EDerwent and MDerwent. As a result, 474409 patents related to communication technology in recent 24 years were retrieved.

2.2 Data Sources and Pre-processing of Communication Industrial Technology Standard

Technology standard is the competition focus of the entire communication industry. There is a great difference among 2G, 3G and 4G in the process of technological development of communication industry. Standard development divided industrial development into obviously different stages. The 3rd Generation Partnership Project (3GPP) is devoted to establishing standard which is globally applicable during the technological evolution from GSM core network to mobile communication of the next generation. 3GPP progresses with a series of constantly modified versions of technology standard (Release) which almost cover most standardization work of 2G-4G in the communicational development process. In order to manage patents concerning technology standard better, European Telecommunication Standards Institute (ETSI) stipulated that it should be informed by the members of the essential patents that its standard concerned as soon as possible according to FRAND principle within a reasonable scope. Technically, in the case of no patent infringement, if a patent is unable to meet the standard requirement during the production, sale, lease, repair and operation of facilities and acquisition of methods, then the patent is identified as an essential patent(Zeng and Peng 2008). In order to increase the transparency of intellectual property rights it has recorded, ETSI has established a specialized database to make regular update and disclosure of essential patents. Enabled in March 2011, ETSI IPR Online Database contains all the standard patents of the 3GPP standardization work from Release 99 until today and provides valuable samples for us to deeply analyze the relationship of industry technology standard and industrial technology development.

Standard patents of communication industrial technology were obtained from ETSI IPR Online Database. By January 2014, we have downloaded all the 2553 3GPP standardization projects from Release 1997 to Release 13 and 50110 essential patents in that database. As for the 50110 essential patents, we made a data cleaning, which included unification of format, deduplication and deletion of informationally incomplete patents, and merged the patent families according to the first patent number marked in each patent data thus finally getting 2208 patent families. Then a second retrieval in DII database based on essential patent number was made to establish a standard patent database of communication technology for a further analysis.
2.3 Methodology

It was reported by the World Intellectual Property Organization that patent was the greatest information source of technology in the world and that it covered more than 90% of global R&D output in which about 70% of the inventions had never been published in non-patent literature (Liu and Yang 2008). For the sake of a clear analysis of industrial technology knowledge, we need to obtain those technology contents from the carrier of technological knowledge. In patent texts, technical terms are the main representation of technological knowledge. Since there is not a structured storage of those technical terms in patent texts, a detailed analysis of the patent texts is thus needed.

First of all, the extraction of technical terms was realized by natural language processing to construct a dictionary of industrial patent content. Compared with journal articles, patent literature lacks of keyword fields so that “technological terminology” needs to be selected from texts as technical terms. The technical terms of patent literature were extracted by two procedures, part-of-speech tagging and rule definition of word phrase extraction. This paper employed Stanford Log-linear Part-Of-Speech Tagger (Toutanova and Manning 2000) developed by NLP group of Stanford University to tag the part of speech of each single word of patent titles and abstracts followed by the extraction of technical terms by rule definition (Liu and Chen 2013). By the way, the accuracy can be above 95% when tagging English words with that system.

Next step was a manual denoising of the extracted technical terms, mainly including the unification of number, the merging of synonyms, the use of en dash “-”, full name and its abbreviation as well as proper phrases. Some commonly used words such as “communication”, “network” and “system” in communication field were deleted and some synonyms such as “D2D” and “device to device”, “OFDM” and “Orthogonal Frequency Division Multiplexing” were merged. After several rounds of manual screenings, 7173 “groups” of word phrases (such as merged synonyms, singular& plural words and abbreviations, etc.) were preliminarily kept as a dictionary of patented technologies in communication field.

Secondly, an analysis of high-frequency words, burst terms and newly emerging terms was undertaken with patentmetrical methods. The statistical analysis of word frequency enabled us to generally estimate technological hotspots and identify the newly emerging technologies (Hu, Zhang et al. 2013). Analyze word frequency of the annual patented technology and high-frequency technical terms were regarded as technological hotspots of that year. The calculation of high-frequency technical terms, burst terms and newly emerging terms is shown as follows:

High-frequency words of Year $i$ $T_i = \text{top 50} \{ F(i,j) \}$, this represents the top 50 most frequent words of one year;

Newly emerging words of Year $i$

$N_i = \{ W(j) \mid F(n,j) = 0, F(i,j) > 0, N > i > 1, i > n > 0 \}$

Burst terms of Year $i$

$B_i = \{ W(j) \mid | F(i,j) - F(i-1,j) | > 10, N > i > 1 \}$

$Y(i)$ represents Year $i$, $N >= i > 0$;

$W(j)$ represents the jth word phrase, $M > j > 0$;

$F(i,j)$ represents the frequency of occurrence of word phrase $W(j)$ in Year $(i)$.
Finally, with the use of vector space modal and cosine similarity algorithm contained in text mining technology, the similarity of patented technology among different years was measured. Vector space model (Wu and Xu 2003) simplifies text content processing to vector operation in vector space and it presents the semantic similarity intuitively via spatial similarity. If documents are represented as the vectors of documental space, the similarity between documents then can be measured by calculating the similarity between those vectors. The computational formula of similarity between vector A and vector B is shown as follows:

\[
\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}
\]

The smaller the distance between vectors is, the more technical terms overlap and more similar the technology content of different years are. If the technology content of different years are very dissimilar, the distance between vectors will be larger, which is accordingly taken as the division point of technological cycle.

3 The Relation Judgment of Leading Enterprises’ Technology Development and Industrial Technology Development

3.1 The Selection of Leading Enterprises in Communication Industry

This paper selected patent leaders of the industry mainly based on the technological strength reflected by patents. Aimed at exploring enterprises’ overall strengths of patented technology, this paper made a comprehensive consideration of the existing evaluation indicators of patent in view of scientific algorithms, available and operable data. Indicators such as Number of Patents, Patent Growth Index and Patent Impact Index were chosen to reflect enterprises’ technological capability in patent.

The technological capability of a company can be preliminarily revealed by the number of patents, but number of patents only is imprecise for it does not take the value of each patent into account. Hall (Hall, Jaffe et al. 2005) and his colleagues had found that market value has a closer relationship with weighted number of patent by citation than with number of patents itself. Therefore, this paper further studied the Top 100 companies with most number of patents per annum about the times cited of their patents so as to give more weight to companies whose patents are frequently cited. Meanwhile, in order to stand out the recent dynamic changes of domainial potential, this paper also added Recent Growth Index to reflect whether patents of institution tend to act actively or slowly. More details about indicators and computing methods can be seen as follows:

Number of Patents: \( P_{ij} \)

Number of patents is the most basic indicator based on which many other indicators are computed. The development tendency of a particular industry can be discovered through an analysis of authorized patents in a certain period. \( P_{ij} \) represents the number of patents company i possesses in Year j.

Patent Impact Index: \( \text{PII}_{ij} \)

Patent impact index reflects the influence of institutional patents on technological development of a field. A lot of researches have indicated that there is a strong positive correlation between patent citation and the importance of technology (Narin, Noma et al. 1987; Albert, Avery et al. 1991; Breitzman and Narin 2001). Patent impact index is the ratio of citations per paper of
all the patents the institution has authorized and obtained during the last 5 years to that of all patents in technology field of the same period, in which the former one has been standardized and get rid of the impact of industry and scale. A PII greater than 1 means that the patents of that institution are frequently cited to a level higher than the average level of the domain. It is thus speculated that the company has a better patent quality because its technologies exert a greater influence on industrial technology development. The computational formula of PII is shown as follows:

$$\text{PII}_{ij} = \frac{C_{ij}}{\sum_{l} C_{lj} / \sum_{l} K_{ij}}$$

(2)

$C_{ij}$ and $K_{ij}$ represent respectively the times cited and the total number of the patents that company $i$ has authorized in $j$ field during the last 5 years.

Patent Growth Index: $G_{ij}$

As a proportion of the number of patent an institution authorized in a particular year (say 2013) to the average amount of patents it has authorized in the last 5 years (2009-2013), patent growth index reflects the patent activity of an institution in recent years. A $G_{ij}$ greater than 1 means an increase of the institution’s patent activity compared with the last 5 years; a $G_{ij}$ less than 1, in contrast, reflects a decrease of its patent activity in recent years. Since $G_{ij}$ is a regulatory index, we set an upper limit of 2 for it in case that companies which do not act actively early in patents possess extremely high $G_{ij}$.

$$G_{ij} = \frac{P_{ij}}{\sum_{l \neq j} P_{lj}}$$

(3)

The three indicators constitute the technological strength (TS) of a company. The technological strength of company $i$ in $j$ field can be calculated by the following formula:

$$\text{TS}_{ij} = P_{ij} \times G_{ij} \times \text{PII}_{ij}$$

(4)

However, the classification methods in this paper also have some limitations. For example, all the patent citations are treated equally without drawing a distinction among the value of them. Furthermore, different reference patterns present unequivalent patent values (Chen, Lin et al. 2007). Although high-value patents such as patents cited by important patentees and beyond this technological domain or of different legal status are given a higher weight, delicate properties like that are not taken into consideration. In general, this paper holds that patent impact reflected by the number of patents and citations could generally represent the competitive situation among companies in a particular field and that a further study of the dynamic changes of leading companies’ patent development could be achieved with consideration of recent growth index.

Based on the downloaded patent data of communication industry from 1990 to 2013, a preliminary ranking and selection of enterprises according to their number of patents was made. After that, a further selection according to patent quality and growth indexes was made to figure out that there were totally 16 enterprises whose patents are among Top 5 in communication industry every year from 1990 to 2013.

Table 1 Patente Codes of 16 Enterprises and Their Representative Companies

<table>
<thead>
<tr>
<th>Patentee Codes</th>
<th>Enterprise Group</th>
<th>Representative Companies</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIDE</td>
<td>NEC Corporation</td>
<td>NIPPON TELEGRAPH&amp;TELEPHONE CORP</td>
</tr>
<tr>
<td>AMTT</td>
<td>American Telephone &amp;</td>
<td>AT&amp;T INTELLECTUAL PROPERTY I LP</td>
</tr>
</tbody>
</table>
3.2 Content Analysis of Patent Leaders’ Patented Technology in Communication Industry

After extracting and collecting high-frequency technical terms of the 16 enterprises’ annual patents, their annual technological hotspots were recognized. On the basis of the patent technology development of communication industry as a whole, a dictionary of patent technology for communication industry was built and the part-of-speech tagging, matching and extraction of technical terms contained in the titles of those enterprises were made. Lately, we made a description and analysis of the enterprises’ annual high-frequency technologies after selecting and ranking each company’s annual technical terms in the light of year and discussed the technological development of the 16 companies leading in communication patents during the 24 years.

It was found that high-frequency technical terms of a company presented an obvious similarity among neighboring years (regular technology period) while changed a lot at one or several points (abnormal technology period). However, during years following the technical terms tended to be similar again. Therefore, we attempted to divide the technical content cycle of each leading company according to the following methods: firstly, sequence all the technical terms contained in the dictionary of communication industry technology according to the order in dictionary of patented technology of communication industry; then counted the word frequency of technical terms occurring in every leading company in different years and made a zero-padding processing for those did not appear in dictionaries; finally made a hierarchical clustering of word frequency variation after standardizing all the technical terms. For a company, the technological emphasis is different in different years, which could be presented by the distribution of technical terms. A hierarchical cluster analysis by SPSS allows us to see more clearly the periodic changing tendency of a company’s technologies.
3.3 Analysis of Communicational Patent Leaders’ Participation in Standardization

A company participates in the formulation of industrial standards to a degree influenced by its internal technological capability (Leiponen and Bar 2008), meanwhile, a higher level of participation also makes it more likely for organizations to transform their technological superiority into global standards thus establishing a de facto standard globally (Bekkers and West 2009). Taking part in the formulation of technology standard is the best way to affect the choice of predominant standard in the future (Bousquet, Fomin et al. 2011). In fact, the development of different versions of Release reflects the development and changes of the patented technology standard of communication industry. The next sector is an analysis of the participation degree of leading companies in every Release during different periods in accordance with their participation in work planning.

Analysis of Patent Leaders’ Participation in Standardization in Terms of the Quantity of Involved Projects

Firstly, this paper drew all the assignee names (enterprises) that had participated in formulating standard for Release 7-13, and then made a second screening of the 16 communicational leaders based on 7 groups of data and finally analyzed the industrial leaders’ participation degree in terms of the number of standardization projects they have participated in. There are many projects undertaken by several companies jointly in standardization work, in that
case, the number of projects will be distributed in proportion to enterprises according to their ranking of participation degree. The proportion of patent leaders’ participation in 3GPP standardization work for Release 7-13 is displayed as the following Table 2:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TELF</td>
<td>0.19008</td>
<td>0.14286</td>
<td>0.19903</td>
<td>0.1734</td>
<td>0.2222</td>
<td>0.12953</td>
<td>0.10204</td>
</tr>
<tr>
<td>HUAW</td>
<td>0.0503</td>
<td>0.10817</td>
<td>0.08252</td>
<td>0.1169</td>
<td>0.1266</td>
<td>0.12953</td>
<td>0.06803</td>
</tr>
<tr>
<td>LUCE</td>
<td>0.07396</td>
<td>0.07933</td>
<td>0.1068</td>
<td>0.1048</td>
<td>0.1111</td>
<td>0.07254</td>
<td>0.05442</td>
</tr>
<tr>
<td>QCOM</td>
<td>0.02071</td>
<td>0.03365</td>
<td>0.08738</td>
<td>0.129</td>
<td>0.0491</td>
<td>0.05181</td>
<td>0.04082</td>
</tr>
<tr>
<td>OYNO</td>
<td>0.1213</td>
<td>0.02404</td>
<td>0.01942</td>
<td>0.0202</td>
<td>0.0207</td>
<td>0.01425</td>
<td></td>
</tr>
<tr>
<td>ZTEC</td>
<td>0.00296</td>
<td>0.01442</td>
<td>0.01942</td>
<td>0.0605</td>
<td>0.0439</td>
<td>0.02979</td>
<td>0.02041</td>
</tr>
<tr>
<td>SMSU</td>
<td>0.02367</td>
<td>0.0024</td>
<td>0.03398</td>
<td>0.0161</td>
<td>0.031</td>
<td>0.00777</td>
<td>0.05442</td>
</tr>
<tr>
<td>AMTIT</td>
<td>0.1213</td>
<td>0.02404</td>
<td>0.01942</td>
<td>0.0202</td>
<td>0.0207</td>
<td>0.01425</td>
<td></td>
</tr>
<tr>
<td>NIDE</td>
<td>0.02071</td>
<td>0.01202</td>
<td>0.02427</td>
<td>0.0282</td>
<td>0.0258</td>
<td>0.00668</td>
<td>0.01361</td>
</tr>
<tr>
<td>MOTI</td>
<td>0.03846</td>
<td>0.04087</td>
<td>0.02427</td>
<td>0.004</td>
<td>0.0078</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLDS</td>
<td>0.01183</td>
<td>0.01202</td>
<td>0.02427</td>
<td>0.004</td>
<td>0.0026</td>
<td>0.01813</td>
<td></td>
</tr>
<tr>
<td>CISC</td>
<td>0.00518</td>
<td>0.0068</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Proportion</td>
<td>55.40%</td>
<td>47.94%</td>
<td>60.19%</td>
<td>66.53%</td>
<td>69.51%</td>
<td>51.94%</td>
<td>36.05%</td>
</tr>
</tbody>
</table>

As we can see from Table 2, the distribution of 3GPP standardization work conforms to the Pareto Principle, that is to say, a minority of enterprises take charge of the majority of projects while most enterprises take charge of a few projects or even not participate in the standard establishment at all. There are more than 300 company members in 3GPP, but only 58 of them have taken part in one or several projects of standardization work. The degree of participation is influenced by the internal technological capability of the company, and the weight of voting when participated in standardization is related to the company’s revenue in communication field (Leiponen and Bar 2008). Table 2 shows that 11 of the 16 leading enterprises have participated in formulating standards of R7-R13 at all. They mainly bear over 50% of the standardization work in Release 7-12 including a proportion of more than 60% in Release 9-11 standardization. Since Release 13 is still under way, the distribution among those enterprises has not been determined yet.

Information of Patent Leaders that Possess Standard Patents

The analysis of standard patents that leading enterprises possess could be seen from Table 3. As is shown in Table 3, there are 9 companies declaring to possess standard patents that are essential to carried standard in the 16 leading companies of communication industry. From Release 4 (2001) to Release 11 (2012), those 9 companies occupied more than 50% of the total number of standard patent in most years except for a slightly lower proportion in Release 4 and Release 10. Almost all the 9 enterprises have participated in standard establishment except LUCE and APPY.
Table 3 Proportions of the Leading Enterprises’ Standard Patents in Total Ones

<table>
<thead>
<tr>
<th>Release4</th>
<th>Release5</th>
<th>Release6</th>
<th>Release7</th>
</tr>
</thead>
<tbody>
<tr>
<td>OYNO</td>
<td>0.351</td>
<td>OYNO</td>
<td>0.277</td>
</tr>
<tr>
<td>QCOM</td>
<td>0.029</td>
<td>SMSU</td>
<td>0.162</td>
</tr>
<tr>
<td>MOTI</td>
<td>0.025</td>
<td>HUAW</td>
<td>0.0777</td>
</tr>
<tr>
<td>APPY</td>
<td>0.005</td>
<td>QCOM</td>
<td>0.049</td>
</tr>
<tr>
<td>NIDE</td>
<td>0.005</td>
<td>TELF</td>
<td>0.0317</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NIDE</td>
<td>0.025</td>
</tr>
<tr>
<td><strong>Total Proportion</strong></td>
<td><strong>35%</strong></td>
<td><strong>51.50%</strong></td>
<td><strong>62.80%</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Release8</th>
<th>Release9</th>
<th>Release10</th>
<th>Release11</th>
</tr>
</thead>
<tbody>
<tr>
<td>OYNO</td>
<td>0.365188</td>
<td>TELF</td>
<td>0.257</td>
</tr>
<tr>
<td>TELF</td>
<td>0.102389</td>
<td>OYNO</td>
<td>0.193</td>
</tr>
<tr>
<td>NIDE</td>
<td>0.068259</td>
<td>NIDE</td>
<td>0.096</td>
</tr>
<tr>
<td>SMSU</td>
<td>0.010239</td>
<td>LUCE</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUAW</td>
<td>0.002</td>
</tr>
<tr>
<td><strong>Total Proportion</strong></td>
<td><strong>54.60%</strong></td>
<td><strong>55.10%</strong></td>
<td><strong>27.50%</strong></td>
</tr>
</tbody>
</table>

4 The Role of Leading Enterprises in Industrial Technology Development

What part the leading enterprises play in the formation of technological paradigms was analyzed from two aspects, the patented technology content of the leading enterprises and their role in standardization of patented technology. Firstly, the role of leading enterprises in forming technological paradigms was defined by the similarity between their key patented technologies in each year and the contemporaneous patented technology hotspots of the whole industry. Specifically speaking, match the yearly high-frequency technical terms from communicational patented technology with the dictionary of communication industry technology and made a zero-padding processing for technical terms that do not appear in that year so that the industrial patents of each year can be represented as a feature vector consisting of 7147 technical terms of different frequency. Make an identical processing for the high-frequency technical terms of the 16 leading enterprises’ patent and we got a feature vector, consisting of 7147 technical terms, which represented the annual patents of a company. We got a correlation coefficient between the technical terms of enterprises and that of industry in that year by means of a correlation analysis. At the same time, after ordering the technical terms of both industry and leading enterprises in accordance with word frequency, we obtained those of high-frequency of which we made a comparison of the ‘Top 20 so as to analyse the relationship between companies’ key technologies and industrial technology hotspots from a qualitative perspective. How the analysis worked is displayed by Figure 3.
4.1 Leading Enterprises Provide Important Technological Base for the Formation of New Technological Paradigm

Comparing the key patented technologies of 16 leading enterprises from 1990 to 2013 with the patent hotspots of communication industry in corresponding year allowed us to learn about the variation trend of the correlation coefficient between patent content of the enterprises and that of industry during 24 years. Put the variation trend mentioned before together with that of the patented technology ranking of enterprises respectively and we got the results as shown below:
In general, the similarity of leading enterprises’ key patented technologies and the patented technology hotspots of the whole industry is basically consistent with the enterprises’ patent rankings. After further studying how the correlation coefficient between 16 industry leaders’ key patented technologies and communicational patented technology hotspots relates to enterprises’ ranking scores, it was found that for 15 enterprises, the consistency of their key patented
technologies and the industrial patented technology hotspots correlate significantly with their rankings in a moderate or higher level. In other words, there is a strong similarity between the patented technology content of 15 enterprises and the industry during the period when they led while a relatively weaker similarity when they did not. That means that leading enterprises’ patent content composes an important part of the industrial patent content and that compared with others leading enterprises’ patented technologies are more similar with that of the industry, which indicates in a degree that the patented technologies of leading enterprises have a significant influence on the formation of industrial technological paradigms.

In view of the high-frequency technical terms of leading enterprises, SMSU, for example, have involved most of the 1999-2008 technology hotspots of the communication industry since 2000. Its development of hotspot technologies is many years ahead of the industry especially for technologies about OFDM, safety and personal communication. The emphasis of SMSU technologies also changed since 2010, though technologies such as LTE, NFC (Near Field Communication), intelligence and portable terminals were consistent with leading industrial technologies. Besides, QCOM started to develop technologies like OFDM, content data and Mobile Terminal from 2000 to 2008, which was also ahead of the industrial technology content and similar with pioneering industrial technologies of that period. After 2009, QCOM changed its technological key points, in which technologies concerning LTE, NFC, WLAN (Wireless Local Area Network) and AP (Access Point) were consistent with leading industrial technologies of that time. As for APPY, since its entry into communication field in 2004-2008, it put the emphasis of technological development on digital camera module, cloud computing, touch screen, multimedia, portable electronic devices, music player, media storage, auxiliary equipment and Bluetooth rather than contemporaneous leading industrial technologies. Some technologies such as smart phones had not become leading industrial technologies until 2009 and others like cloud computing even not appeared as pioneering technologies until 2009.

It can be concluded from the above analysis that technologies of leading enterprises is ahead of the leading industrial technologies of the same period and compose an important part of contemporaneous technologies, thus providing an important technological base for the formation of a new technological paradigm.

4.2 Leading Enterprises Promote the Formation of Industrial Technological Paradigm by Technology Standardization

In view of the similarity between leading enterprises’ patented technologies and the industrial technologies of standard patent, enterprises that have participated in standard establishment generally have a higher level of similarity with industrial technical standard than enterprises that have not. The correlation coefficient between such leading enterprises as CANO, IBMC and LUCE that have not involved in standard establishment and standard techniques is far lower than that for leading enterprises which have involved in. This indicates to some degree that it is likely that leading enterprises have blended their own key technologies into the new generation of industrial technology standards during the participation in standard establishment so as to affect the formation of technological paradigm. The process can be displayed by Figure 5.
5 Conclusion

Since the combination of technology and standard is becoming more and more close, patent leaders tend to play an increasingly important role in promoting the development of industrial technology. With a study of the function mechanism of leading enterprises’ technology on industrial technology development, it is found that leading enterprises provide significant technological base for the formation of new technological paradigms and promotes the formation of industrial technological paradigm by technology standardization, which makes them a powerful driving force for industrial adjustment and upgrade as well as innovative development.

(1) Technologies of leading enterprises are ahead of leading industrial technologies of the same period and compose an important part of the contemporaneous pioneering technology thus providing important technological base for the formation of new technological paradigms.

(2) Leading enterprises of communication industry play a significant role in promoting the formation of both industrial technological standard and technological paradigms via active participation in standardization work and a large amount of standard patents in the process of industrial standardization.

(3) Leading enterprises blend their own key technologies into industrial technological standard so that they determine partial content of the industrial technological paradigms. After technological paradigms are formed, leading enterprises achieve a transformation from pioneering technology to leading technology, which makes the patent content of communication industry leaders an important part of leading industrial patented technologies.

The conclusions above are all drawn based on theoretical hypothesis and empirical studies of communication industry. The rule of industrial development this paper reveals is applicable to communication industry. At the same time, this research is expected to provide a research paradigm for technological analysis of other industries.

Acknowledgments

This research is supported by National Natural Science Foundation of China (grant number 71503030), China Postdoctoral Science Foundation (2015M571312)
References


Network Community-based Technological Cooperation Identification

Lu Huang 1 Wen Miao 2 Yi Zhang 3 Huizhu Yu 4 Kangrui Wang 5

1,huanglu628@163.com
Beijing Institute of Technology (China)

2,a2689813532@126.com
Beijing Institute of Technology (China)

3,yi.zhang@uts.edu.au
Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney (Australia)

4,675244410@qq.com
Beijing Institute of Technology (China)

5,kangruiwbb@163.com
Beijing Institute of Technology (China)

Abstract
Technological cooperation becomes necessary in extending technological application and improving market competitiveness, and thus, it is important to identify collaborative technologies that meet strategic purposes. This paper aims to construct a technological network based on the core terms derived from patent documents to help technology owners achieve such target. A four-step analytic framework is constructed, including: 1) community division, which allows the selection of cooperative technology to be controlled within a more relevant technical range; 2) indicator analysis that can be used to understand the technological situation within the community from three main aspects, i.e., degree, clustering coefficient and line weight; 3) the purpose of cooperation, which is an important basis for the selection of cooperative technology; and 4) technology selection, which is to combine cooperation purposes and indicator analysis to select technology. Finally, a case study on China’s artificial intelligence (AI) technology is conducted to demonstrate the feasibility of this method, and the findings will help decision making in related AI fields.

Conference Topic
Patent analysis; Social network analysis; Mapping and visualization;

Introduction
With the continuous development of science and technology, the degree of knowledge specialization is getting higher and higher. To improve the ability of technological innovation and market competitiveness, technological cooperation has received increasing attention (Tu, Mohler & Ma, 2017; Kotsemir, Kuznetsova & Nasybulina, 2016). In addition, it’s hard for an individual enterprise to bear high risk and huge capital investment requirement in the process of new product development, and technological cooperation can reduce the adverse impact of failure on enterprises.

With advantages in value-added information, structured data format and low acquisition cost, patent has been widely used to analyse modern technologies, such as technical information flows, technological trends (Segev & Kantola, 2012), technological innovation (Lee & Kim, 2010), and technological strategy (Ernst, 2003). Bibliometrics have been commonly used in patent analysis due to its simplicity and convenience (Bellotti, Kronegger & Guadalupi, 2016;
Landini, Malerba & Mavilia, 2015; Chen & Fang, 2014), but the shortage is that it heavily depends on bibliographic information, which doesn’t include detailed technology information (Yoon & Park, 2004). As a remedy, a content-based approach fills the deficiency by analysing detailed information extracted from text, and keyword-based analysis (KWA) is a representative method of content-based approaches. The combination of KWA and network models (Huang, Zhu & Guo, 2014; Stermitzke, Bartkowski & Schramm, 2008), known as keyword-based network, can directly reveal the relationships between keywords and help to analyse technologies within the theoretical framework of networks (Wu, 2016; Zhang, Shang & Huang, 2016; Choi & Hwang, 2014), and the engagement of advanced information technologies (e.g., machine learning) further enhances such ability in relationship identification (Zhang, Zhang & Zhu, 2016).

Community structure is a feature of networks. Node in a network can be divided into groups, and nodes in a group are tightly connected while node connections between different groups are sparse. Here group is called as community. Community in actual systems has significance meanings, and can help solve many actual problems. In social network, community may be based on human occupation, age or other factors (Girvan & Newman, 2001); in citation network, community may be divided according to research areas (Redner, 1998); in World Wide Web, different communities may represent different themes on web pages (Flake, Lawrence & Giles, 2002). Under this circumstance, indicator analysis becomes an important way to understand the topology of networks, and some common indicators include degree, density, compactness, line weight, closeness, clustering coefficient etc.

This paper proposes a method that identifies technological cooperation based on network community structure. Community division allows the selection of cooperative technologies to be controlled within a more relevant technical range and makes the selection result more accurate. We identify technological cooperation purposes into three categories, i.e., extending technology applications, improving technology level and identifying possible technological connection. Network indicator analysis is adopted to select cooperative technology. Different purposes need different indicators. This paper mainly conduct the following research: community division, indicator analysis, purpose of cooperation and technology selection. Finally, a case of China’s artificial intelligence technology is studied to illustrate the availability of the proposed method.

This paper is organized as follows: In the next section, we introduce the methodological framework, including patent collection, network construction, community analysis and identification of technological cooperation. A case study of China’s artificial intelligence technology is discussed in the following section to demonstrate the feasibility of our method. Finally, we draw some conclusions and discuss the future study.

Method

The proposed method is comprised of four steps: patent collection, network construction, community analysis and identification of technological cooperation. The process is showed in Figure1.
**Patent Collection**

This step aims to propose a search strategy and collect patents. To make a good search strategy, some necessary factors should be considered, including the features of patent databases, actual research purpose and time span. For example, special patent database does not reveal enough information about the legal status, the same family, the patent.

**Network Construction**

Before we construct keyword-based patent networks, keyword acquisition is required, i.e., extracting keywords from patent abstracts, filtering invalid words (such as number, academic terms and other basic words) and selecting a number of keywords as the research objects. According to a co-occurrence matrix, we construct a keyword-based network, where each node represents a keyword and the line represents the co-occurrence relationship between the two end nodes.

**Community Analysis**

Based on the network constructed in the above step, we divide the whole network into some communities based on the connection tightness between nodes, allows the selection of cooperative technology to be controlled within a more relevant technical range. Then, we adopt network indicators (e.g., degree, clustering coefficient and line weight) to explore each community, providing theoretical basis for technological cooperation identification.

1) Community Division

Intuitively speaking, community refers to a set of nodes, and the connections within a community are dense while the connections between different communities are sparse. Network community indicates a set of nodes that have some common attributes or some similar effects. The algorithm for community division is as follows.

\[
Q = \frac{1}{2m} \sum_{i,j} A_{ij} \left( k_i k_j \right) \delta(c_i, c_j), \quad \delta(c_i, c_j) = \begin{cases} 1, & c_i = c_j \\ 0, & c_i \neq c_j \end{cases}, \quad m = \frac{1}{2} \sum_{i,j} A_{ij}, \quad 1 \leq i,j \leq n \quad (1)
\]

Where \(A_{ij}\) is the line weight between node i and node j (1), \(k_i = \sum_j A_{ij}\) denotes the sum of all
the line weights connecting node i, $c_i$ and $c_j$ are the community indexes of node i and node j respectively, $m$ is the sum of all line weights of network and Q represents modularity, whose value is between [0, 1]-the higher Q value the better community division.

In the initial state, every node is a separate community, and the community number is n. Then, communities merge together one by one. The iteration ends when Q gets the maximum value.

2) Indicator Analysis

Before selecting a target technology, we need to understand the technology situation by analyzing nodes connection within the community, such as the influence and connection tightness of nodes, and node cohesion. Indicator which is used to analyze the relationship of nodes can be taken to research the above situations. Considering the diverse implications of indicators in handing diverse technological problems (Yan & Luo, 2016; Jeong, Kim & Choi, 2015; Koseoglu, 2016), we try to summarize certain common indicators and their technological implication in Table 1.
**TABLE1. Detailed network indicators introduction**

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Description</th>
<th>Technology implication</th>
<th>Paper Case</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Density</strong></td>
<td>The tightness of nodes connection in the network</td>
<td>A high value means the network is highly connected, indicating the technology is well developed. While a low value means a sparse network whose development is immature.</td>
<td>(Flake, Lawrence &amp; Giles, 2002)</td>
</tr>
<tr>
<td><strong>Degree</strong></td>
<td>The number of direct links between target node to others</td>
<td>A high value means target node is highly linked to other nodes, indicating the node technology is important.</td>
<td>(Harten &amp; Retél, 2016)</td>
</tr>
<tr>
<td><strong>Component</strong></td>
<td>The number of parts that make up the network</td>
<td>A small value means the network is highly connected, and technologies are closely linked.</td>
<td>(Taibi, Gualberti &amp; Bazilian, 2016)</td>
</tr>
<tr>
<td><strong>Clustering coefficient</strong></td>
<td>The extent of links between neighbour nodes of target node</td>
<td>A high value means the links between neighbour nodes is dense, which indicates the target node surrounded by strong connections is very important. This indicator judges the importance of node by the density of its neighbour nodes’ links.</td>
<td>(Urban, Zhou &amp; Nordensvard, 2015)</td>
</tr>
<tr>
<td><strong>Line weight</strong></td>
<td>The strength of the relationship between the two nodes connected</td>
<td>Representing the intensity of the interaction between individuals, a high value indicates the two technologies are highly relevant.</td>
<td>(Tao &amp; Xue, 2016)</td>
</tr>
<tr>
<td><strong>Closeness</strong></td>
<td>The extent of difficulty from target node to all other nodes in the network</td>
<td>A high value means the information of target node can be easily spread to all other nodes, indicating the target technology has an important influence in the network.</td>
<td>(Lu, 2010)</td>
</tr>
<tr>
<td><strong>Betweenness</strong></td>
<td>The number of the shortest path passing through the node to the total number of the shortest paths.</td>
<td>Reflecting the effect and influence of nodes in the whole network. Used to measure the intermediary and brokerage capability of technology.</td>
<td>(Radicic, Douglas &amp; Pugh, 2015)</td>
</tr>
</tbody>
</table>
In our paper, we select indicators based on cooperation purpose, and different purposes will be analyzed with different indicators. Considering actual requirements on technology management and analysis, we specifically select three indicators in this paper, i.e., degree, line weight and clustering coefficient. Degree reflects the influence of a node, line weight expresses the connection tightness between two nodes and clustering coefficient corresponds to the cohesion of a node. The detail explanation of these indicators is given below.

- **Degree**
  In a network, the degree of one node means the number of direct links between the node and other nodes. High degree value means the node is highly linked with other nodes, indicating the technology reflected by this node is important. The calculation formula (Jeong, Kim & Choi, 2015) is given below.

  \[ d_i = \sum_{j=1}^{n} l_{ij}, \text{ where } i \neq j \text{ and } l_{ij} = \begin{cases} 1 & \text{if } w_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \]

  where \( w_{ij} \) denotes the line weight between node \( i \) and node \( j \), and \( d_i \) denotes the number of lines connected to node \( i \), and can reflect the importance of node in a large level.

- **Clustering coefficient**
  The clustering coefficient of one node denotes the extent of links between the neighbor nodes of a target node. A high value means the links between neighbor nodes is dense, which indicates the target node surrounded by strong connections is very important. This indicator measures the importance of a node by the density of its neighbor nodes' links. The calculation formula (Yan & Luo, 2016) is as follows.

  \[ cc_i = \frac{2l'}{k(k-1)} \]

  Where \( cc_i \) represents clustering coefficient of node \( i \), \( k \) denotes the number of lines connect to node \( i \) directly, \( l' \) denotes the number of lines existed in the \( k \) nodes, and \( k(k - 1)/2 \) denotes the number of possible links between those \( k \) nodes.

- **Line weight**
  Line weight describes the relationship between two nodes. In this paper, the value means the co-occurrence frequency of two technologies in the patent set. The definition of line weight (Barrat, Barthelemy & Pastor-Satorras, 2004) is introduced below.

  \[ w_{ij} = \frac{n_{ij}}{N}, \text{ where } i \neq j \]

  Where \( w_{ij} \) represents the line weight of node \( i \) and node \( j \), \( N \) denotes the patent number of patent set, \( n_{ij} \) denotes the number of patents which contain both technologies of node \( i \) and node \( j \). A high value of \( w_{ij} \) means node \( i \) and node \( j \) are highly related.

These indicators are used to select technological cooperation object, and different technological cooperation purposes need different indicator analysis. Next, we will introduce cooperation purposes.

**Identification of technological cooperation**

This step aims to select target technology based on the results from indicator analysis. To do that, we need firstly figure out the purpose of the cooperation.

1) **Purpose of cooperation**

In this paper, we mainly focus on three types of technological cooperation purposes, and the definitions are given below:

Extend technology applications
• The aim of the cooperation is to extend the application scope of own technology. The ideal cooperative technology should be those that are widely connected with other technologies.

Improve technology level
• Representing cooperate with professional technology to innovate own technology, which can improve the market competitiveness of own technology compared with similar technology.

Identify possible technological connection
• If technology a1 and technology a2 both connect closely with Technology b respectively, there is great possibility that a1 and a2 will connect tightly in the future. So, the cooperation between this two technologies is in line with technology development trends and easy to succeed.

2) Technology selection

Based on the above research of cooperation purposes and indicator analysis, we select different indicators to realize the three cooperation purposes, and realize the selection of cooperative technology.

In a community, the node with high degree value is widely connected with other nodes, which means this technology is easily to build connection with other technologies. Those technologies with high degree are good selection for extending technology applications.

Clustering coefficient indicates the relationship between node and its surrounding nodes. A high value shows strong cohesion, which means the surrounding nodes are highly connected and form a topic group. If one aims to improving technology level, the member technologies of the group are suitable to cooperate.

Line weight measures the connection tightness between two nodes. Technology couple with high value shows the close contact of existing connection, can be used to identify possible technological connection.

The above is the introduction to the method framework. We use indicator analysis to understand the technological information within the community. Then, based on cooperation purpose we select corresponding cooperative technology. Next, we take china’s artificial intelligence technology as a case to demonstrate the availability of the method.

Case study: China’s Artificial Intelligence

To demonstrate the feasibility of our proposed method, china’s artificial intelligence technology is selected as a case. Artificial intelligence (AI), a technical science, aims to research and develop theories, methods, and application system for simulating, extending and expanding human intelligence. The reason we choose China’s artificial intelligence is that China has the world's leading voice and visual recognition technology, and its research ability of artificial intelligence is impressive. Besides, according to two important reports of the White House released in October 2016, Preparing for the Future of Artificial Intelligence and The National Artificial Intelligence Research and Development Strategic Plan, China has surpassed the US in the total number of cited journal articles that relate to deep learning techniques in 2014. What’s more, on March 5, 2017, at the opening ceremony of the National People's Congress, Premier Li Keqiang announced that China would speed up the research and development of new industries such as artificial intelligence. This is the first time that China's highest national conference has incorporated artificial intelligence into government work reports. This shows that artificial intelligence has become the priority of China's economic agenda, and the government has decided to support its growth. China’s artificial
intelligence has indeed made great achievements, and technological cooperation will broaden this field’s future development.

**Data**

To study the development of AI in China, the patent set of Chinese AI technology was collected. The search strategy is “TS= ("artif* intelli*" OR "comput* intelli*" OR "deep learn*" OR "machine learn*" OR "big data" OR "cloud comput*" OR "pattern recogn*" OR "neural network*" OR "data mining*") AND PN= (CN*)”. As a result, 10,229 China’s patents between 2012 and 2016 were acquired from Derwent Innovations Index (DII).

**Result**

We draw a keyword-based network where nodes represent keywords and lines represent the co-occurrence relationships between two end nodes. Conducting community division, we got three communities displayed in Figure 4 (b1, b2 and b3 are the independent part of each community). Next, we combine figures and indicators analyse technology situation of each community.

![Figure 4. Community division result of the technology network.](image)

(a) shows the overall division effect. (b1), (b2) and (b3) are community1, community2 and community3 respectively.
TABLE 2 displays the results of degree and clustering coefficient analysis of three communities.
In degree analysis, we list out the top three technologies of each community. Comparing the degree values with community size, we find that in each community these technologies are almost all connected with other technologies, indicating these technologies are the best selection to cooperate with to extend technology application. From Figure 4-(b1), (b2) and (b3) we can find out the sparse nodes, and expand their technology application. The results are as follows:
- Community 1:
  Automatic control can cooperate with cloud calculation, internet-of-Things (IoT) or wireless sensing technology.
- Community 2:
  Hadoop architecture can cooperate with big Data, video recognition or real time.
- Community 3:
  Sensitivity analysis can select machine learning, neural network or data mining.
In clustering coefficient analysis, the top three technologies of each community are displayed. These technologies are highly professional, can promote own technology to have rapid growth. Improve own technology in the same type of technology in the competitiveness. From Figure 4-(b1), (b2) and (b3) we can identify the potential technological cooperation as follows:
- Community 1:
  Ultrasonic sensor can select to cooperate with automatic control.
- Community 2:
  Dynamic process algorithm can select to cooperate with image identification.
- Community 3:
  Feature selection algorithm can select to cooperate with Hidden markov model algorithm.

<table>
<thead>
<tr>
<th>Partition</th>
<th>Size</th>
<th>Degree</th>
<th>Clustering coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community 1</td>
<td>19</td>
<td>cloud calculation</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>internet-of-Things (IoT)</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wireless sensing technology</td>
<td>17</td>
</tr>
<tr>
<td>Community 2</td>
<td>9</td>
<td>big data</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>video recognition</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>real time</td>
<td>7</td>
</tr>
<tr>
<td>Community 3</td>
<td>22</td>
<td>machine learning</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>neural network</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>data mining</td>
<td>20</td>
</tr>
</tbody>
</table>

TABLE 3 reflects the result of line weight analysis, listing out the top 5 technology couples of each community. Figure 5 is the network diagrams of these technologies.
In community 1, Fig.5-(c1) is the corresponding network diagram of these technologies, where we can see that there are other connections between the six technologies besides the top 5 lines. For example, internet-of-Things (IoT) and wireless communication technology. For both of them are closely connected with cloud calculation, there are great possibility that internet-of-Things (IoT) and wireless communication technology can cooperate with each other and build close connection in future. After a similar analysis, we find in community 2 that pattern recognition and real time is possible close connection in the future and can
cooperate with each other. In community 3 machine learning and fuzzy algorithm, data mining and fuzzy algorithm are potential technology combinations.

### TABLE 3. Top 5 technology couples of line weight in each community

<table>
<thead>
<tr>
<th>Partition</th>
<th>Ranking</th>
<th>Node Couple</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community1</td>
<td>1</td>
<td>cloud calculation—virtual technology</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>cloud calculation—internet-of-things (IoT)</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>cloud calculation—storage module</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>cloud calculation—wireless communication technology</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>cloud calculation—power supply</td>
<td>0.009</td>
</tr>
<tr>
<td>Community2</td>
<td>1</td>
<td>pattern recognition—image identification</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>pattern recognition—video recognition</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>pattern recognition—big data</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>image identification—video recognition</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>big data—real time</td>
<td>0.002</td>
</tr>
<tr>
<td>Community3</td>
<td>1</td>
<td>machine learning—neural network</td>
<td>0.043</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>machine learning—data mining</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>neural network—data mining</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>machine learning—learning algorithm</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>neural network—fuzzy algorithm</td>
<td>0.004</td>
</tr>
</tbody>
</table>

![Figure 5. Top 5 technology couples of line weight in each community.](image)

(c1), (c2) and (c3) correspond to community 1, community 2 and community 3 respectively.

**Discussion and Key Findings**

We constructed networks based on keywords extracted from Chinese AI patents, and divided it into 3 communities. In each community, we took indicator analysis to select cooperation technologies according to different cooperation purposes.

According to the connection tightness between nodes in the network, the network is divided into three communities. The contribution of community division is to refine the area of technical selection to a more relevant scope. We grasp the technical situation within each community by indicator analyse. When selecting cooperation partners, cooperation purposes are considered. This paper summarizes three kinds of cooperation purposes: extend technology applications, improve technology level and identify possible technological connection. Corresponding analysis indicators are: degree, clustering coefficient and line weight. Based on the above research, we can identify the following technological cooperation:

1) Cooperate with high degree technologies to extend technology applications, e.g., automatic control and cloud calculation in Figure 4-(b1).
2) Cooperate with high clustering coefficient technologies to improve technology level, e.g., ultrasonic sensor and automatic control in Figure 4-(b2).
3) Identify possible close technological connection by existing technology connections, e.g., pattern recognition and real time in Figure 5-(c2).

CONCLUSIONS AND FUTURE STUDY

In this paper, we propose a method based on network community to identify technological cooperation, mainly include community division, indicator analysis, purpose of cooperation and technology selection. This article is characterized by using community division and combining with technology purposes. Community division filtered out unrelated technologies and made the results more accurate. Different technology purposes adopt different network indicator analyses, according to the analysis results selecting technical cooperation object. Finally, we select China’s artificial intelligence technology as a case to demonstrate the feasibility of our method. This paper has following contributions: proposing a method to identify technological cooperation and benefit experts of technical cooperation identification. There are some limitations on our research. For example, in this paper, the network is divided into three communities. Applying other community division algorithm may get different results. Besides, this paper adopt three indicators based on research purposes, there are some other indicators can be used to analyse network topology, such as density, closeness and so on. Those indicators may have other new findings. In future, more accurate community division algorithm and more diverse indicator analysis should get further research.

Acknowledgments
We acknowledge support from the Basic Research Foundation of Beijing Institute of Technology (Grant No. 20152142010), and the National Science Foundation of China Young Funds (Grant No. 71103015).

References:
Koseoglu, M. A. (2016). Mapping the institutional collaboration network of strategic management


Abstract
The major challenges of topic visualization in texts are to enhance, interpret, and explain the visualization results. Two strategies for enhancing and interpreting results of underlying topic visualization in texts are proposed. In order to overcome the limitations of visual display space capacity, the open coding method is introduced to assign terms to different categories, with each category corresponding to a visual space. In view of the difficulties in interpretability and understandability of the visualization results, this paper investigates the characteristics of context dependencies of the underlying topic and its terms, and puts forward a three-layer context model, which includes three levels, that is, domain context, topical context and linguistic context. The ideas and methods for constructing the context model in this study can be extended to other areas, such as text segmentation, automatic summarization and other text mining and knowledge discovery.

Conference Topic
Mapping and visualization

Introduction
Underlying topics can reveal main content of a text set. If underlying topics can be extracted and represented in the visual space, people can quickly obtain the key content of a text set. The hidden knowledge structure, mode and potential regular features in texts can be found and deep knowledge discovery can be realized.

A term is the smallest semantic unit in texts and clustering terms in visual space is an efficient way to visualize underlying topics. Due to the inner relationship among terms in texts in a specific domain, those terms expressing the same or similar meaning will show close spatial neighborhood relations in a 3D visual space. However, there are two difficulties in finding topics by the method of term clustering in the visual space at present:

The first is the limitation of visual spaces (Zhang & Zhao, 2013). Due to limitations of visual spaces (computer screen size), it is impossible to put all terms and underlying topics in the same space to present to the users. Conversely, if all the underlying topics and their terms are integrated in one visual space, advantages of the visualization method is weakened or even lost.

The second is the complexity and ambiguity when the visualization results are explained (Zhang et al., 2014). The underlying topic visualization is realized by using the method of transposed-vector-space to represent terms. But in the process of converting texts into matrix form, a lot of useful semantic information is ignored. The terms themselves are ambiguous in their meanings, which increase the complexity in interpretation of visualization results. Therefore, it is necessary to find a qualitative method to ensure correct interpretation and in-depth analysis of visualization results.

This paper aims to propose hybrid strategies to overcome those two difficulties. Research questions include: (1) how to overcome the limitation of a visual space that can only display limited number of terms? (2) how to reveal the true meaning of underlying topics and its terms profoundly?

Findings of this study can be used to: (1) reduce the number and size of objects in the visual space; (2) help interpret and understand the results of underlying topic visualization.
Related work
Topic visualization by clustering terms in visual spaces is used a lot in research on knowledge discovery in texts (Zhang, Zhao & Dimitroff, 2014; Zhang & Zhao, 2013; Zhang & Wolfram, 2009; Zhang, Wolfram, & Wang, 2009; Zhang et al., 2008). It is recognized that the limitation of visual spaces and the difficulty to interpret the visualization results are major problems and not addressed well until now (Zhang, Zhao & Dimitroff, 2014; Zhang & Zhao, 2013).

Methods
Two strategies for enhancing and interpreting results of underlying topic visualization in texts are proposed. One is to introduce coding technics into the process of underlying topic visualization. The other is to interpret the visualization results under a three-layer context model. Those two strategies are implied in underlying topic visualization of public company prospects.

Data sources
According to Corporation Law, Securities Act and Interim Measures for the Administration of Initial Public Offering and Listing on the Growth Enterprise Market, the issuer of initial public offering (IPO) must disclose all factors that may directly or indirectly cause significant adverse effects on operating conditions, financial conditions, sustained profitability and growth of an issuer. The relevant documents require the issuer to fully, accurately and specifically describe the relevant risk factors according to its actual situation. The issuer shall make a quantitative analysis of the risk factors disclosed and make a targeted qualitative description for those factors that cannot be made with quantitative analysis.
According to the Industry Classification Guidance for Public companies issued by China Securities Regulatory Commission on April 4, 2001, this study selected the computer application and service industry (classification code G87) as an sample for the analysis and all verbal contents in the chapter of “Risk Factors” in the prospectus of public companies in the computer application and service industry form the text set to be analyzed.

Underlying topic visualization
Texts were firstly represented as a term-document matrix and each item in the matrix is the term frequency. Then, a Multi-dimensional Scale analysis was executed to clustering terms in the visual space following the transition of the term-document matrix to similarity matrix. The software used here is SPSS of version 22.

Extending visual spaces by using coding technics
Aimed at addressing the first difficulty of underlying topic visualization, the open coding method is introduced to control the number of objects in the visual interface and realize the multi-layered visualization.
It is also necessary to design corresponding strategies of coding according to different situations while the coding technics are introduced into the process of underlying topic visualization (Zhang & Zhao, 2013). There are two coding approaches for text sets with different characteristics:
The first approach is to code terms after word segmentation. This strategy is applicable to completely unstructured free texts, like text sets composed of messages in online Q&A or other social media, chatting records in the instant messaging and others. These texts don’t have a clear text structure and often have no titles and even paragraphs. For this strategy, the object of open coding is terms and the task is to divide the terms obtained from the segmentation into several categories.
The second approach is to code texts before word segmentation. This strategy is applicable to text sets composed of weakly structured official texts, such as business plan, prospectus of public companies and other business or technical texts. This type of texts has large textual length, tight structure and clearly defined sub-title.

The first approach requires the deep participation of experts and the workload of coding is bigger; the second one just needs a small number of manual coding due to the use of the topic information of the text, and the workload of the coding is smaller. The advantages of introducing coding technics into underlying topic visualization are to control the number of terms and topics in the visual interface, solve the problem that the visual interface can only present limited objects and ensure the visual effect (Zhang & Zhao, 2013). At the same time, open coding is a classification environment and its disadvantage is that it is possible to divide the terms in the same underlying topic into different categories artificially.

**Interpreting visualization results under a three-layers context model**

After the underlying topic visualization for the topic related text sets, the underlying topic and its terms have been obtained. Only by taking full account of the underlying topic and the context of the term can we correctly understand the true meaning of them (Dey, 2000). On the one hand, the meaning of a term could be ambiguous, such as “Defense”, which can be a term expressed in the sports field, but also can expressed in the military field (Haghighi et al., 2009). Another example is “Flexibility”. For the separate analysis of this term, we do not know whether it refers to the elasticity in physics or flexibility in economics (Mei & Zhai, 2006). On the other hand, when the object of a term is different, its meaning will be different, such as “Size”. It can be the size of the enterprise, market size or staff size. Therefore, the underlying topic or terms must be interpreted in the context where they occur.

We proposed a three-layer context model include domain context, topical context, and linguistic context, which will be shown in the section of results.

**Results**

*Results of open coding of sub-titles and text under those sub-titles*

In this study, the theme of the target text set is risk factors of public companies in the computer application and service industry and the purpose of open coding is to divide the risk factors into several categories. Prospectus is a kind of business texts that is with complete format, and has clear sub-titles marking the corresponding fragment theme and representing the central content of the corresponding fragment. All section titles form the central theme. Therefore, the second approach of coding is adopted. The first step is to extract the sub-titles of all risk factors and form codes; the second step is to classify codes and name them; the third step is to form a category table.

A schema with ten categories were generated by coding and the ten categories are market risk, operational risk, financial risk, products and technology risk, investment project risk, internal management risk, inter-control risk, human resources risk, industry risk, and political risk. The original text is divided into many subsets by coding and terms in each sub text set can be visualized, clustered and displayed separately to achieve the purpose of controlling the number of objects in the visual interface.

*Results of underlying topic visualization*

Texts in the category of operational risk were selected to be visualized and analyzed. The results of underlying topics visualization are shown in Figure 1 and three underlying topics were found by term clustering.
Interpretation of the visualization results under the framework of context model

The result is interpreted under the framework of the three-layer context model which is shown in Figure 2.

The first layer is domain context. The scope of domain context firstly depends on the determination of research objectives, namely discovery of text knowledge of public companies; and then comes selection of the study object. This paper selects the public companies in the computer application and service industry as the object of empirical study. Public companies have many types of texts, such as prospectus, announcements, annual reports and others. Here prospectus is chosen as the target text set. At this point, the follow-up studies should consider issues under the context of the prospectus of public companies in the computer application and service industry.

The second layer is topical context. Prospectus is large in length and has many modules. The sections of risk factors in prospectuses of public companies in the computer application and service industry are selected to form the target text set. The first approach of coding is adopted here and ten categories were generated by coding. Terms included in each category will be clustered into a visual space. In our experiment, the category of operational-risk is the topical context for the analysis of visualization results.
The third layer is linguistic context. Terms in the same underlying topic and terms in the different underlying topics in the same visual space constitute the linguistic context. If there is a need for further research, you can return to the original text to locate a core term and get more profound conclusions through manual reading, analysis and comparison, generalization and summarization and other methods. Further explanation according to the linguistic context is given below:

Topic 1 primarily reflects the internal risk factors. One example is the company’s business concentration risk. Most companies in computer application services industry always adopt a “product specialization” style to focus on advantages of their products. But it also brings the over-concentration operational risk.

Topic 2 mainly reflects the seasonal risk. In China, government departments, the military, medium-sized enterprises, operators and other customers usually adopt budget management system and centralized procurement system. Generally, the project budget is approved in the first half year, tendering, procurement and construction process are executed in the second half year. The user market demand typically peaks in the second half year. Revenues are lower in the first half, and cash inflow from operating activities are mainly concentrated in the second half, while the R&D investment, staff salaries and expenses across months in a year are of equilibrium. This together leads to imbalance and unevenness in company’s profits and cash flows generated from operating activities.

Topic 3 mainly reflects risks from external environment. The first part is the risk of the instability in supply of raw materials and suppliers. The second part is the constraints of the business license qualification.
Discussion and Conclusion

In order to overcome the limitation of visual spaces, the open coding method is introduced to assign terms into different categories, with each category corresponding to a visual space. A large number of objects are divided into different visual spaces for display, which overcomes the space limitation and ensures the effect of visualization.

Aimed at the shortcomings in the interpretability and understandability of visualization results, this paper proposed a three-layer context model, including domain context, topical context and linguistic context. At the same time, this study also found that the first two contexts have been defined when determining the research objectives and data sources. The third context can be improved and enhanced in the research design part.

From the perspective of domain context, if the target text set is the business text set or technology text set, the corresponding strategies should be designed based on these characteristics. Because this kind of texts mostly has clear structure, titles and sub-titles representing topic content. If the target text set is the completely unformatted network health information text, such as questions and answers about diabetes on the online Q&A website, another coding strategy should be constructed. From the point of view of the topical context, the understanding of underlying topics shall be carried out under the subordinate major topic. In the view of linguistic context, more contextual information should be provided for target terms.

One limitation of this paper is the bias of its data source. This paper only uses formal Chinese text (prospectus) with more normative and rigorous text structure to verify the two strategies of open coding and the context model and doesn’t test free texts. In future work, more experiments will be performed on free text sets, such as text sets composed of message in online Q&A.

Acknowledgements

This research was supported by the National Natural Science Foundation of China under Grant 71403190 and 71420107026. It is also supported by Research Fund for Academic Team of Young Scholars at Wuhan University (Big Data and Business Analytics, Whu2016013).

References


Abstract
In this study we tried to find a new way based on LDA to detect topics of projects and the papers produced by the projects to grasp the hot topics and topic evolution of the funding agency. We detected 5 hot topics of the projects on ocean acidification supported by NSF and articles produced by these projects based on LDA by computing subject-intensity. And we analyzed the evolution of topics of projects on ocean acidification supported by NSF. After comparing the topics we detected from projects and related papers by LDA with the recommendations proposed by the Ocean Carbon and Biogeochemistry Group, we find the topics detected by LDA is practical and effective. And some suggestions to detect topics of projects and articles with LDA are put forward.

1 Introduction

Topics of projects supported by national funding agency show research frontier and hot research topics of the field. Articles supported by the projects are presentation of the research progress. Some researchers like to find the status and hot topics by browsing the related projects and reading the papers supported by the projects. How to grasp the hot topics and the evolution of the research topics from the information of projects and their research output is becoming a critical task for researchers and funding agency. With the rapid growth of digital information, how to get the hot topics and grasp the tendency quickly and save time to direct to the related topics from large scale information is useful.

In this article, firstly, we try to find a new way based on LDA to detect hot topics of the projects supported by NSF and the papers written by the projects, analyze the research topic evolution of the projects supported by NSF. Then we compare the result of our experiment with some authoritative view of some professional broad to
test the validity of our method. At last, we make some suggestion to detect topics of projects and articles with LDA.

2 Literature Review

2.1 Analysis of projects supported by research Funding agency

According to metric analysis of scientific fund and their outputs, many researchers focus on statistics analysis of allocation of scientific projects (e.g. J.Ying, 2003), financial support (e.g. V.A. Markusova, 2010), subject field (e.g. D.Z. Zhao, 2010; G. Lewison & G. Dawson, 1998; S. Eckhousea, et al., 2008), regional and organizational allocation (e.g. L. Leydesdorf & C. Wagner, 2009; X.W. Wang, et al., 2012; V.A. Markusova, 2010). Some researchers carried out performance analysis of research funding by number of articles, number of citations (e.g. X.W. Wang, et al., 2010; P. Zhou, et al., 2012). These researches payed much attention to metadata indices and external characters of projects and articles to evaluate the performance of the projects. It’s hard to explore the research topic and evolution of research topic. Hot topics analysis and emerging topics detection have been an new and challenging field in scientometrics, C.W. Belter (2012) analyzed the research filed of articles supported by NOAA, to identify the topics of articles supported by the funding with co-words and bibliographic coupling. Few of studies relates to topic detection and evolution analysis of research projects with text-mining technology or LDA.

2.2 Detection topics with Latent Dirichlet Allocation (LDA)

Normally, identification and detection of topic are the basic of research frontier detection. K.K. Bun and M. Ishizuka (2001) used TF*PDF (Term Frequency * Proportional Document Frequency). H.X. Liu, et al. (2010) extracted topics with TF*IDF in their emerging trend detection. This method extracted words according to word frequency and syntax structure rather than the semantics of the topics. T. Hofmann (1999) proposed PLSI (Probabilistic Latent Semantic Indexing) as the first topic model. D. Blei, et al. (2003) proposed Latent Dirichlet Allocation (LDA), which is a generative model based on statistical foundations and which also tends to produce clearly interpretable output in a more robust manner. Several related, augmented variants of LDA have been developed in recent years (D. Hall, et al., 2011). These and other methods built on the foundations of LDA have come to be known as “topic models”. One key strength of LDA comparing with common document clustering methods based on term frequency or K-Means is that LDA explicitly represents each document as being generated from a document-specific mixture of multiple topics (a multinomial over words), which present somewhat semantic context, rather than single word. This flexibility tends to lead to more clear and interpretable topics. The models have generated significant interest because they create fine-grained, immediately interpretable topics that are robust against synonymy and polysemy. In
addition, generating topics and document topic assignments require virtually no human supervision (A. McCallum, et al., 2006).

L. He and F. Li (2012) extracted the topics of scientific documents with topic model, after computing the intensity and impact of the topics, they carried out research trend analysis. H. David (2008) mapped the changing history of ideas using topic models. Stanford Topic Modeling Toolbox (R. Daniel and R. Evan, 2009) can extract topics of documents, according to the map of documents and topics, time slices, theme of intensity change trend can be presented.

3 Data and Method

3.1 Procedure of extracting topics from document with LDA

In this study, we use LDA1 to extract topics of research projects. LDA posits that each document is generated as a mixture of topics where the continuous-valued mixture proportions are distributed as a latent Dirichlet random variable, and every topic is composed by a group of words with different probability. LDA get topics of the documents by iterative estimate. Given a document collection with D documents, assuming the document collection contain W words, K topics, the procedure of extracting topics from document with LDA is as below:

1. For each document d ∈ D, according to Dirichlet distribution \( \theta_d \sim \text{Dir}(\alpha) \), we get topic distribution parameter of document d, \( \theta_d \)

2. For each topic z ∈ K, according to Dirichlet distribution \( \phi_z \sim \text{Dir}(\beta) \), we get polynomial distribution \( \phi_z \) of topic z

3. For ith word of document D, wd,j, according to polynomial distribution zd,j~Mult( \( \theta_d \) ), we get topic zd,j, according to polynomial distribution wd,j~Mult ( \( \phi_z \) ), we get word wd,j

According to the procedure mentioned above, the model cannot get Dirichlet distribution \( \theta_d \) and \( \phi_z \) directly. In general we get the approximate parameter values with the method of parameter estimation, normally with Gibbs sampling algorithm based on Markova chain(Q.Z. Yao, et al., 2011; P.J. Zhang & L. Song, 2012). The value of \( \alpha \) and \( \beta \) should be predefined empirically, \( \alpha = 50/K \) (K means the number of topics of the collection of documents) and \( \beta = 0.01 \) is thought appropriate in experiment(Z.Z. Wang, et al., 2013; X.P. Zhang, et al., 2010).

LDA is a kind of unsupervised topic mining algorithm. D.M. Blei, et al. (2003) computed the most appropriate number of topics with perplexity when he put forward LDA, T.L. Griffiths(2004) used Bayesian method, some other researchers determined the number of topics of LDA empirically (e.g. J.P. Cao, et al., 2014; B. Wang, et al., 2015). All these studies evaluated the model from different views.

1 http://nlp.stanford.edu/software/tmt/tmt-0.2
3.2 Topic similarity

Topic similarity represents the degree of correlation between two different topics, generally using KL distance to measure the differences of two themes in different distribution (K. Cui, 2010). The representation of the word distribution $\phi_{k1}, \phi_{k2}$ of two different topics is as follows:

$$D(P(w|k1)||P(w|k2)) = \sum_{w=W} P(w|k1) \log \frac{P(w|k1)}{P(w|k2)}$$

The standard KL distance is non-symmetric. If the value is 0, it represents the two topics are exactly the same, actually the degree of correlation of the contents between the two topics should be symmetrical. So, we deform the KL distance in order to measure the differences of different topics’ distribution in the symmetric KL distance measure: $KL_s(A,B) = (D(A||B) + D(B||A))/2$.

According to symmetrical KL distance relationships, we calculate the degree of correlation between the topics of project and the topic of fund papers. Then we get the correlation $\psi_{k1,k2}$, $K1$ and $K2$ express different topics. This could be used to calculate the evolution of the topics of the project papers.

3.3 Topic intensity

Topic intensity reflects the popularity of the topic in the document, and its evolution can be measured by observing the changes of the topic over time. This paper adopted the method in literature of P. Wang (2011). According to the distribution of document topics, we give a weight to each document. The document which focus on fewer topics gets a higher score, and the one which has broad contents gets a lower score. Entropy is a good standard that can be used to calculate document weight. Finally, we combine with the weights of the document and the polynomial distribution of document’s topics to calculate the topic intensity.

$$\text{entropy}(d_m) = - \sum_{K=1}^K \theta_{m,K} \log \theta_{m,K},$$

$$w_m = 1 - \frac{\text{entropy}(d_m)}{\max\{\text{entropy}(d_1), \ldots, \text{entropy}(d_M)\}},$$

$$T_t(z_k) = \frac{\sum_{m=1}^{M_t} w_m \theta_{m,k}}{\sum_{m=1}^{M_t} w_m}.$$

$m,d$ represents the probability of the document $m$ belongs to the topic $K$, $M_t$ represents the number of documents within the period of time $t$, $T_t(z_k)$ represents the intensity of the topic $K$ in the period of time $t$. Through Slicing the topic intensity of document by years, both the topic concentration of the projects and the scope and variation trend of projects and papers supported by the projects can be observed. Combining with the topic similarities between the NSF projects and the paper supported by them, we can get the topic intensity of the papers supported by NSF fund in different topics:

$$T(z_k) = \frac{\sum_{m=1}^{M} \sum_{K'=1}^{K'1} \theta_{m,K'} \psi_{k,k'}}{M}$$
3.4 Data and preparation

We selected projects information funded by United States NSF and papers supported by the projects from Web of Sciences on ocean acidification as the experiment data. Ocean acidification is a phenomena which due to the absorption of excess carbon dioxide in the air caused the increase of ocean acidity or the decrease of PH. The current global ocean is in the fastest speed of acidification since 5500 million years. A paper published in Nature firstly discussed the problem which caused widespread concern in the scientific community (K. Caldeira, 2003), and ocean acidification became the hotspot of many international conference, such as the international symposium “Ocean under the conditions of high concentration carbon dioxide” (J.A. Raven, 2005) and so on. NSF is an important funding agency in the United States, which concerns on public health, food safety, water quality, environment, education and some other livelihood issues, and ocean acidification is one of the important funded branches with program code 1382. Considering the representativeness and practical significance, we collected the data on ocean acidification and detect the hot topics and the topic evolution in this field interested by NSF.

Firstly, we retrieved the projects with the search “((ocean* OR seawater* OR marin*) AND (acidif* OR (carbon* OR Nitro* OR sulf*) AND PH))” in the title and abstract of the awards funded by NSF. Secondly, we got all of the awards data containing program code 1382. Then, we merged the two parts and remove the duplicates as the first set. Meanwhile, we get the related papers supported by the funded projects with the similar search from Web of Sciences as the second set. The data instruction is shown in Table 1. The abstract information of funding awards contains the contents of the research topic, the research significance, research members and institutions, and broader impact and so on. Here we interested in research topics, therefore, it is necessary to clean data. This study reference the cleaning rules in publication of T. Chen et al. (2015), and remove the information which useless to the research topic. After artificial cleaning, it generates some independent text documents, then we clear the stop words and stem them, afterward we use LDA model mining the handled data. Refer to LDA model, it has to iterate 1000 times at most to get the best groups as the topics mining result. Then, we compare the different numbers of topics and determine the conspicuous groups to get the suitable number of topics with the method mentioned in publication of B. Wang et al. (2012).

<table>
<thead>
<tr>
<th>ID</th>
<th>Data source</th>
<th>Year range</th>
<th>Data Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NSF</td>
<td>2004-2016</td>
<td>357</td>
</tr>
<tr>
<td>2</td>
<td>WoS</td>
<td>2008-2016</td>
<td>655</td>
</tr>
</tbody>
</table>

Table 1 two sets for LDA modeling

The data is retrieved in April 2016
4 Result

4.1 NSF projects and funded papers of ISI

As shown in Figure 1 which pictures the trend of the quality of NSF projects which divided into three groups: standard grant, continuing grant and fellowship and the total amount of funding from 2004 to 2016. The number of projects in 2004 and 2005 was still relatively small, but had increased significantly from 2006 to 2008 when nearly 20 projects were funded every year, until 2010 reached the highest. The number of projects in 2011 is the double of that in 2008, then NSF paid a high attention on ocean acidification in the next years and the trend keeps ever-increasing. As to funding types, the proportion of fellowship is increasing, and the number of standard grant kept 30 around every year, while continuing grant varied considerably relatively. The total amount of funding is higher in 2009, 2010, and 2012 than in the other years when it maintained the fluctuations in the level of $20,000,000. Abnormally, it had reached unprecedented levels in 2009 when the new U. S. President Barack Obama took over officially and the Federal Ocean acidifying Research AND Monitoring Act firstly began to implement in March. The implement of the act not only was the result of ocean acidification becoming more and more serious, but also made it highly concentrated.

Figure 2 shows the trend of papers of Web of Sciences funded by NSF from 2008 to 2016. It can be seen that the overall number is gradually increasing year by year except 2012 when the production may be resulted in the cut of funding and the number of projects in 2011. Above analysis roughly reflect the changes in projects and papers funded by the NSF organization each year, which is unable to figure out what the projects funded by NSF exactly focused, hence, the further study on the topics of the projects content is necessary.

Figure 1 the number tends of projects and funding of NSF
4.2 The intensity of NSF projects

We got 5 topics on ocean acidification based on the LDA topic model calculated and parameter optimization. Then, we pick up the top 15 keywords of each topic, abandoned the words which appear in three groups simultaneously and have no represented meaning, such as ocean, carbon and so on, finally show in Table 2.

<table>
<thead>
<tr>
<th>ID</th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>oxid</td>
<td>measur</td>
<td>water</td>
<td>respons</td>
<td>develop</td>
</tr>
<tr>
<td>2</td>
<td>activ</td>
<td>sampl</td>
<td>isotop</td>
<td>coral</td>
<td>model</td>
</tr>
<tr>
<td>3</td>
<td>organ</td>
<td>dissolv</td>
<td>atmospher</td>
<td>effect</td>
<td>ecosystem</td>
</tr>
<tr>
<td>4</td>
<td>microbi</td>
<td>inorgan</td>
<td>surfac</td>
<td>condit</td>
<td>commun</td>
</tr>
<tr>
<td>5</td>
<td>reaction</td>
<td>total</td>
<td>observ</td>
<td>level</td>
<td>studi</td>
</tr>
<tr>
<td>6</td>
<td>abund</td>
<td>seawat</td>
<td>sea</td>
<td>Speci</td>
<td>climat</td>
</tr>
<tr>
<td>7</td>
<td>form</td>
<td>determin</td>
<td>record</td>
<td>temperatur</td>
<td>global</td>
</tr>
<tr>
<td>8</td>
<td>cell</td>
<td>water</td>
<td>season</td>
<td>reef</td>
<td>biolog</td>
</tr>
<tr>
<td>9</td>
<td>sourc</td>
<td>method</td>
<td>sediment</td>
<td>experpi</td>
<td>process</td>
</tr>
<tr>
<td>10</td>
<td>repres</td>
<td>situ</td>
<td>cycl</td>
<td>organ</td>
<td>marin</td>
</tr>
<tr>
<td>11</td>
<td>environ</td>
<td>paramet</td>
<td>rate</td>
<td>affect</td>
<td>data</td>
</tr>
<tr>
<td>12</td>
<td>metal</td>
<td>alkanin</td>
<td>estim</td>
<td>predict</td>
<td>result</td>
</tr>
<tr>
<td>13</td>
<td>composit</td>
<td>test</td>
<td>variabl</td>
<td>marin</td>
<td>scale</td>
</tr>
<tr>
<td>14</td>
<td>fluid</td>
<td>develop</td>
<td>dissolut</td>
<td>physiolog</td>
<td>approach</td>
</tr>
<tr>
<td>15</td>
<td>function</td>
<td>ion</td>
<td>global</td>
<td>calcif</td>
<td>identifi</td>
</tr>
</tbody>
</table>

According to the words groups of topic1, the main key words, including organ(ic), microbi (ology), reaction, cell, function, etc. After consult with some experts, we identify topic 1 is about the relationship between ocean acidification and the marine organic materials and the general marine biology. The concentration of seawater carbon is a crucial environmental parameter to the marine organisms. Some of the marine organisms need calcium carbonate in seawater to build shells and skeletons of
raw materials and many need keep the concentration of carbon dioxide in the extracellular fluid and water in a certain range in order to ensure the normal respiratory function. Acidification directly affects the sea creatures living conditions, due to diversely different biological effects, this topic focuses on the large scale observation and investigation of marine organisms and microorganisms, which can directly reflect the severe impacts of ocean acidification.

Topic 2 has the main keywords of dissolve(e), inorgan(ic), paramet(er), sample, alkalin(e), ion(ic), which basic covering the cue words of carbon dioxide measurement system elements, so is it as the topic2’s main theme. The measurement parameters of the ocean carbon dioxide system generally includes four factors: dissolved inorganic carbon, alkalinity, pH value, the equilibrium fugacity of carbon dioxide in water samples (A.G. Dickson, 1993; E.E. Bockmon, 2014). Changes of carbon dioxide content in the ocean influence the degree of ocean acidification, and the measurement of carbon dioxide system can evaluate the survival state of sea creatures, and the observation on biological activities in different carbon dioxide environment can predict the effects to marine biosphere caused by ocean acidification, and the experiment measurement on sample can provide practical solutions. Measurement of ocean carbon dioxide system is the beginning and important process of the research of ocean acidification. Without the measurement of carbon dioxide system there would be no accurate data sources for study and detection of the experiments, so do the research findings lose some persuasive.

Considering the main keywords in topic 3, such as sea water, isotop(e), observ(e), sediment, dissolut(e), atmosphere, it can be estimated topic 3 is about the application of isotope in the study of relationship between ocean acidification and climate change rate and global cycle of carbon and nitrogen, and at the same time, this topic contains the effects of the sedimentary and dissolved calcium carbonate to ocean acidification. The main theme of topic 4 is on the impacts of ocean acidification on calcification of marine creatures, which be inferred by the words of coral, reef, effect, affect, calcification etc.. It’s known that low pH value will affect the rate of calcification of marine organisms. As one of the most important livings in ocean, coral reef contains diverse organisms and has remarkable economic benefits. Ocean acidification does not only influence the activities of calcified organisms, and indirectly influence the flora and the fauna depending on them (O. Hoegh-Guldberg, et al., 2007; J.M. Pandolfi, et al., 2011). Besides coral reef, there are some other creatures with calcification, such as mollusks, echinoderms, fishes and so on. The responses of different species to ocean acidification are so different, and the study is beneficial and necessary to both economic and biological systems.

The subject of topic 5 is easily to be concluded by the words of model, ecosystem, climate, biolo(gy), data and so on which represent biological model construction, analysis and comparative of data model, the relationship and influence between ocean acidification and global ecosystem. Constructing model is an important means and method to study the complicated problems. There are a wide variety of biological species in the marine system and the different areas of ocean are not identical. Through the construction of marine biogeochemical model, data analysis and
comparisons of the similarities and differences of various species in different ocean zones (M.C. Leal, et al., 2015; G.G. Waldbusser, et al., 2013), more precise judgments and predictions on the rise of carbon dioxide caused by climate change will be made.

As discussed above, topic 2 is about data acquisition, topic 5 is the methods on studying the problem, and topic 1, 3, and 4 are about the different impacts of ocean acidification caused and causing, which were the habitats of marine organism, the economic benefits and climate change.

In order to test if our experiment detect the right hot topics of ocean acidification, we compared the 5 topics detected by LDA with 11 recommendations on ocean acidification research direction of the United States over the next 5-10 years proposed by the Ocean Carbon and Biogeochemistry Group(2009), established the ocean acidification subcommittee and chaired the ocean acidification research direction of the United States over the next 5-10 years. We found a certain corresponding relationship in the research content can be found. For example, topic 4 focuses on the effects of ocean acidification on calcification research, especially on the coral reef, while the 2rd and 5th recommendations are also advocated this topic. Topic 5 deliberates the data model on ocean acidification which also be seen in the 8th recommendation. Therefore, it can be considered that the topic clustering results of LDA model is a good way to express the main research contents of ocean acidification projects NSF funded.

According to the topic intensity formula above to calculate and draw the topic intensity variation of the projects funded by NSF, as illustrated in Figure 3.

As shown in the line graph, topics 2 and 5 are facing the decline trends overall and the downward trend of topic 2 is more obvious. In 2004 when the intensity ranked second, however it dropped to the bottom after 2009, indicating that the attention of NSF on the measurement of carbon dioxide system is weakening. The relatively smaller decline of topic 5 suggest that the construction of the model to simulate the response of marine ecology to ocean acidification is gradually reduced, meanwhile, the continuously high proportion indicting that topic 5 was still an important parts of NSF
projects. In contrast, before 2006 the strength of topic 4 was generally in a lower level, then gradually increased after 2006 and maintained at a high level in recent years. It suggests that the study on marine calcifying organisms to ocean acidification is a hot research topic of NSF funded nowadays. Although in several years topic 1 and 3 have abnormal performance, but the intensity changed little in the overall sight. Topic 1 reduced slightly in recent years, while topic 3 remained stable relatively, indicating that topic 3 is still under development. In the sight of the fluctuations of the intensity change, the volatility of the 5 topics presents large amplitude in the earlier years and gradually became small, which implies that NSF has been clear and steady on the direction of main topics of ocean acidification.

4.3 The intensity of papers supported by NSF of ISI

LDA model is used to mine the topic of the papers supported by the selected NSF supported of projects in Web of Sciences. After calculating the similarity with NSF projects, the intensity variation of the papers on the above 5 topics by years is shown in Figure 4.

![Figure 4 the intensity variation of the topics papers of ISI funded by NSF](image)

Observing the intensity variation of each topic, topic 2 and 5 stay on a low level in the most of the years and changed little, reflecting that papers related to these topics are relatively fewer. The output of topic 2 ranked in second in 2008 when it was principal areas of concern, but subsequently, it has been always in a low level. Topic 1 has been gradually declining. Combined with the changes in the number of papers every year, the output of topic 1 kept in a certain amount but not in an intensity scale. At the same time, it is apparently found that the intensity of topic 3 has varied apparently, it slowly raises from 2008 to 2012 and declines rapidly in 2013, then firms up a little in 2014 and 2015. In contrast, topic 4 remained sluggish growth in a certain extent, and it could be concluded that topic 4 outputted the highest proportion’s core papers.

Comparing the intensity of both projects and papers supported by NSF in WOS, the variation of papers’ intensity is consistent with the projects’. When the hot topics of some projects changed, the intensity of the papers’ topic followed. The intensity rank of topic 5 in the papers is lower than that in the projects, which represented that topic
5 had a relatively low yields, while the other topics is in the almost same rank indicates that the intensity of other topics in projects is consistent with that in papers.

5 Discussion and Conclusions

In this study, we apply LDA topic model to mine the main topics on the projects funded by NSF of ocean acidification. We also calculate the variation of the intensity of each topics by years, which shown the hotspots NSF paid attention to and the proportion arranged every year. Then we use the WOS papers about ocean acidification funded by NSF to measure the publication output of each topics and found some consistence with the funded awards.

The method used in this study has opened up a new perspective to analysis the funded award data, which not only reveals the hot topics of a certain subject, but also presents the trend of the topics evolution. The results are able to predict the proportion and funding trends in the future, and provide some references and supports for making plans or adopting policies.

Several suggestions in using LDA to detect hot topic and topic evolution are put forward. First, In the process of data, it is necessary to clean up and to eliminate the general stop words to exclude the useless information. Secondly, in the process of topic explanation, it is important to consult some experts or to make some preparation for the professional knowledge.

Some ongoing study have been carry out to compare the topics with projects in the same field supported by NSF and NSFC, that means to carry out experiment on Chinese with LDA to detect hot topics and topic evolution.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No: 71373260).

References and Citations


Cui, K. The Research and Implementation of Topic Evolution Based on LDA [M], National University of Defence Technology, 2010.


Bundle of sleeping-beauties:  
The case of Paul Hagenmüller and solid-state chemistry

Adil El Aichouchi1  Philippe Gorry2

1 adil.el-aichouchi@etu.u-bordeaux-montaigne.fr  
SPH EA 4574, Bordeaux Montaigne University, (France).

2 philippe.gorry@u-bordeaux.fr  
GREThA UMR CNRS 5113, University of Bordeaux, (France).

Introduction

Paul Hagenmüller, and solid-state chemistry
Paul Hagenmüller (1921 – 2017) was one of the most well-known and influential French chemists of the second half of the 20th century. He is considered the father of solid-state chemistry (SSC) in France and head of a school at the University of Bordeaux. Solid-state chemistry, also referred to as materials chemistry seeks to synthesize new solid phase material and to study them by both chemical and physical methods (Hagenmüller, 1999; Teissier, 2010).

Sleeping beauty in scientific literature
In scientometrics, the phenomenon of delayed recognition has been well described since the pioneering observations of Garfield, and referred to today as sleeping beauties (SB). Van Raan (2004) defined SB as articles that go unnoticed (“sleeps”) for a long time and then, suddenly, receives a lot of citations by a “prince” (another article). Scientometrics offers three kinds of quantitative criteria: average-based criteria, quartile-based criteria and parameter-free criteria (Ke, 2015).

Objectives
Studying delayed recognition can be very useful to understand the dynamics of scientific change. The present work aims to study to what extend the phenomenon of delayed recognition have affected the birth, dissemination and autonomy of solid-state chemistry. We analyze the scientific production of Hagemuller’s body of work and the citation life of his publications looking for sleeping beauty. Then, we attempt to identify the relevant prince, and try to understand the reason for delayed recognition and awakening mechanisms.

Materials & Methods
To collect the publications of Paul Hagenmüller, we used Scopus® database. A collection of 646 publications and 13,676 citations were harvested until the 31st December 2015. Excel was used for statistics and calculations. Using Ke et al’s criteria (2015), the beauty coefficient (B) was calculated for all of Hagenmüller’s papers in order to detect top SB. The quantitative analysis was complemented by a historical analysis of SSC.

Results

Hagenmüller published his first paper in 1964 in Journal of Inorganic and Nuclear Chemistry. During his 38 years’ career, he published 620 original articles, and collaborated with more than 150 different authors. He published his last paper in 2002. His scientific work totaled 13,676 citations and reached 651 citations/year in 2015.

The calculation of the B coefficient using Ke’s equation for all of Hagenmuller’s papers reveals the existence of possibly 7 SB published between year 1965 and 1985 (Table 1), and awakened between year 1993 and 2010 (Figure 1 and 2).

<table>
<thead>
<tr>
<th>SB</th>
<th>Authors</th>
<th>Year</th>
<th>Source</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Delmas C., et al.</td>
<td>1980</td>
<td>Physica B+C</td>
<td>307.9</td>
</tr>
<tr>
<td>7</td>
<td>Kasper J.S., et al.</td>
<td>1965</td>
<td>Science</td>
<td>236.52</td>
</tr>
</tbody>
</table>

Table 1. Top Hagenmuller delayed papers.

These publications are related to 3 different research subjects: SB #1, 2, 4, 5, and 6 are related to sodium-ion batteries; SB#3 is related to sodium cobaltate and SB#7 to silicon clathrates. Citation histories of SB #1, 2, 4, 5, and 6 are aggregated in Figure 1, and in Figure 2 for SB#3 and 7. In both figures, reference line lt (black line); distance dt maximizing the awakening time (dashed line) and, awakening time (vertical line) are reported. Moreover, the publications trend per year of the related research subject is described for sodium-ion
batteries (Figure 1; black dotted line) and for sodium cobaltate and silicon clathrate research fields (Figure 2; dotted line).

**Figure 1.** Citation history of SB 1, 2, 4, 5, and 6 and publication trends.

**Figure 2.** Citation history of SB 3 and 7 and publication trends.

**Discussion**

**Sleeping-beauties dormancy and awakening**

The change in social situation and development of the battery technology might explain the awakening of first group of SB. Studies on Na-ion batteries have attracted significant renewed interests since 2010 because they are considered safer and are a lower-cost candidate for large-scale energy storage compared to lithium-ion. The same hypothesis can be formulated for SB#3 on sodium-cobalt oxides which is related to research on positive electrodes for lithium and sodium electrochemical. Studying clathrate-type phases of silicon and related elements were considered as simple crystallographic curiosities for a long time. The discovery of fullerenes renewed the interest in them (Pouchard & Cros, 2014), and one might postulate that SB#7 awakening was stimulated by the Nobel prize in chemistry awarded in 1996. We must stress that SB #3, 4 and 6 were published in French. Since Hagenmüller’s laboratory transitioned slowly from French to English as it has become the main language for scientific communication (Table 2), it is possible that the diffusion of those papers was limited by this choice.

**Table 2.** Hagenmüller’s articles language use

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Publications in French</td>
<td>76%</td>
<td>53%</td>
<td>16%</td>
<td>2%</td>
</tr>
<tr>
<td>Publications in English</td>
<td>14%</td>
<td>43%</td>
<td>83%</td>
<td>98%</td>
</tr>
</tbody>
</table>

**Bundle of sleeping-beauties**

SB cluster have been described by Van Raan (2016) through co-occurrence word and concept map. Ke et al. (2015) also observed that papers belonging to the same topic exhibit remarkably similar citation histories. In parallel, Prince clusters’ and the need of multiple PRs to awake SB have been discussed by Teixeira (2016). In our case study, we would like to introduce the concept of “bundle” of SB as a cluster of SB related to one research field or an author, not necessarily through co-citations. Since SB is potentially associated to resistance to discovery and science paradigm shift, we might postulate that it would affect not only isolated publications but also related work from the same author or the same research field for which a consolidated B coefficient could be calculated based on the cumulated SB citations to ascertain the delayed recognition for the whole body of work (B coefficient 325,741).

**References**


Web of Science™ as a Research Dataset

Katy Börner,1,2 Valentin Pentchev,2 Matthew Hutchinson,2 James Pringle,3 Jason Rollins,3 Yadu N. Babuji,4 & Eamon Duede5

1katy@indiana.edu
School of Informatics and Computing & Network Science Institute, Indiana University, USA
2maahutch@iu.edu & vpentche@iu.edu
Network Science Institute, Indiana University, USA
3jason.rollins@clarivate.com & james.pringle@clarivate.com
Clarivate Analytics, USA
4yadunand@uchicago.edu
Computation Institute, Knowledge Lab, University of Chicago, USA
5eduede@uchicago.edu
Computation Institute, Knowledge Lab, Committee on the Conceptual and Historical Studies of Science, University of Chicago, USA

Introduction

The Clarivate Analytics Web of Science (WoS) has served as a research dataset for more than 9,000 scholarly articles in the past 15 years alone—across a wide range of fields and disciplines from toxicology to computer science to economics. Scientists and scholars have been particularly interested in the WoS citation network, a massive graph containing billions of links that can proxy the structure and dynamics of not only scholarly communication, but knowledge diffusion, the evolution of fields, and the career lifecycles of individuals and institutions. To power these investigations, scholars are increasingly employing a number of compute-intensive methodologies, sophisticated big data infrastructures, and so called collaborative “discovery science” tools and techniques. Suddenly, in addition to deep, domain specific expertise, world-class computational knowhow appears to be a new prerequisite for analysis of scholarly data at the scale represented by WoS. While cloud-based computing and tools are more prevalent and accessible than ever before, harnessing these technologies remains both a challenge and opportunity for researchers and data providers (i.e., Clarivate Analytics and similar commercial data vendors and non-commercial aggregators). While the opportunities made possible by scholarly data at the size and scope of WoS for discovery and innovation are limited only by imagination, two general prospects come readily to mind. First, access to these data coupled with the appropriate computational and analytical capabilities opens up a wide range of funding and subsequent publishing opportunities in high impact venues. Second, data providers can pursue new business opportunities, including novel data access models, new types of analytic products, and new kinds of academic/industry partnerships. In this poster paper, we briefly explore 1) the new computational infrastructures that are being developed to enable collaborative research that leverages scholarly datasets such as WoS that are both big and proprietary; 2) some recent findings that have been made possible by these infrastructures; and, 3) new commercial offerings that have been enabled and demanded in response to increasing reliance on the WoS as a research dataset.

New Computational Infrastructures

Research leveraging big, scholarly datasets like WoS presents researchers with challenges related to the data’s size, inherently relational format, and sensitive (proprietary) nature. To overcome these challenges, researchers have developed a new generation of enclave supported, high performance, and cloud-based, collaborative research environments that are both big and proprietary; 2) some recent findings that have been made possible by these infrastructures; and, 3) new commercial offerings that have been enabled and demanded in response to increasing reliance on the WoS as a research dataset.

IUNI WoS Data Enclave

The Indiana University Network Science Institute (IUNI) acquired the complete set of Clarivate Analytics’ Web of Science XML raw data (Web of Knowledge version 5). The data was parsed and stored in a well-documented Postgresql raw database, entity-relationship diagram, database schema, and
Leveraging WoS in Research and Practice

Fostering Global Collaboration

Among others, IU started to use the IUNI WoS data to understand existing and foster global research collaborations. The world map in Figure 1 shows the co-affiliations of authors that listed “Indiana Univ” and at least one other non-U.S. institution as affiliation on 1,590 scholarly papers published in 2004-2013. There are 344 affiliation locations (not counting IU) and 641 co-affiliation links. Nodes denote author locations and are area size coded by degree with the exception of IU, which has 1,592 co-affiliation links. Links denote co-affiliations, e.g., an author with three affiliations IU, X, Y will add three links; the two links that connect IU with X and Y are drawn in red while the link between X and Y is given in green. Links are size coded by the number of co-affiliations with the top-three being Beijing, China (155), Eldoret, Kenya (115), and Pushchino, Russia (90).

Cloud Kotta

One platform specifically developed with WoS in mind is Knowledge Lab’s Cloud Kotta (CK). CK is a secure data enclave and analytics platform that serves the research needs of social sciences (Babuji 2016). By hosting CK in the Amazon Web Services cloud, the developers were able to take advantage of virtually limitless compute, cost-effective storage and the ability to implement a fine-grained security model ensuring the authorized collaborators could access both data and compute resources from any where in the world (Babuji 2016). Moreover, CK supports multiuser, rapid ideation and research iteration through a novel Python library that enables specific functions in an analysis code, written in a Jupyter Notebook to be seamlessly and securely submitted to the CK execution fabric (Babuji 2017). By allowing researchers to develop and share analysis code interactively over secure data like WoS, CK has removed the need for deep computational infrastructure expertise. The complete WoS XML dataset was ingested into a relational database housed in CK using a custom parser that has been made freely available on GitHub (see: https://github.com/alexander-belikov/wos_parser). The Cloud Kotta WoS database schema can be found on CK’s documentation pages (see: http://docs.cloudkotta.org/dataguide/wos.html). More about Cloud Kotta can be found at http://docs.cloudkotta.org.

Impact vs. Disruptiveness

Researchers at the University of Chicago’s Knowledge Lab and Northwestern University’s NICO have used WoS data going back to 1900 to study the relationship between team size and impact and the relationship between team size and disruptiveness. This work, currently under review, finds striking differences between the scientific output of large and small teams. Looking across all fields represented in WoS, small teams are shown to disrupt science, patents, and software with new ideas and opportunities, while large teams contribute to existing ones. Figure 2 shows the relationship between impact and disruptiveness of articles (left panel) indexed by WoS, patents (middle), and software (right). In all three spaces, there is a strong, inverse relationship between citations and disruptiveness as team size increases.
New Commercial Offerings

The value of the Web of Science as a search and discovery tool is well established at thousands of research institutions worldwide. But the commercial opportunities for the use of its high-quality metadata outside of the platform for big data studies are still emerging. When researchers need to study broad-scale trends in science, technology, and innovation, they very often turn to the Web of Science as the most comprehensive citation source to provide over 100 years of consistent, global publication data. Increasingly, user requests for this data take the form of custom reports, curated data sub-sets, and large-scale raw XML delivery. Clarivate Analytics is actively looking at compelling ways to meet these customer demands with new commercial products and data delivery choices. These options must balance scale and ease of use, with security and control over access to the proprietary WoS data. The lessons learned in the development of Cloud Kotta and IUNI WoS Data Enclave will very likely be instructive here, as they have proven their utility and leverage a mix of custom code built on proven commercial cloud services. Both self-service data access and secure use of analytical tools in a cloud “sandbox” seem like attractive features of these environments that could make commercial sense to meet the evolving expectation of Web of Science customers.

Acknowledgments

This work as partially supported by and contributes to research for IBM, Facebook, Jump Trading, AWS, Clarivate Analytics, the National Institutes of Health under awards P01 AG039347 and U01CA198934 and the National Science Foundation under awards NCSE-1538763, EAGER 1566393, and NCN CP Supplement 1553044. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work uses Web of Science data by Clarivate Analytics.

References


Developing Armenian Science Citation Index: Obstacles and Challenges

Gzoyan Edita¹ Manukyan Aram² Sargsyan Shushanik³

¹ editagzoyan@gmail.com
Center for Scientific Information Analysis and Monitoring (CSIAM), Institute for Informatics and Automation Problems (IIAP),
1 Paruyr Sevak, 0017 Yerevan, Armenia

² 1.aram.manukyan@gmail.com
Algorithm Languages and Programming Department, Computer Science and Informatics Faculty, National Polytechnic University of Armenia (NPUA)
105 Teryan str., 0009 Yerevan, Armenia

³ shashaniksargsyan8@gmail.com
Center for Scientific Information Analysis and Monitoring (CSIAM), Institute for Informatics and Automation Problems (IIAP),
1 Paruyr Sevak, 0017 Yerevan, Armenia

Introduction

After 70 years under the Soviet rule, in 1991 the Republic of Armenia has regained its independence. Soon, the republic was faced with a number of problems in nearly all areas, among them also in science. A sharp decline of the republic’s once huge scientific potential was marked. As a result of brain-drain and reorientation the scientific personnel involved in the research and development was reduced nearly 3-fold, while science funding was reduced by nearly 90% (Gzoyan et all, 2015). R&D institutions and organizations were rapidly closing or restructuring and reorganizing into business projects due to the brain-drain and lack of finances. The first two decades after the independence were the years of survival and adaptation for the Armenian science and scientific community (World Science Report, 1996; World Science Report, 1998).

Apart from the above-mentioned reasons, the R&D personnel also experienced subjective shortcomings: differences in Soviet and Western type education and methodologies, language difficulties, poorly equipped laboratories, etc. All the mentioned factors were hindering the integration processes in the scientific field.

The mentioned reasons and a general chaotic situation in the country resulted in the localization of the science in Armenia. The Soviet policy of “iron curtain” was preventing collaboration between the Soviet and Western scientists, among which also publications in Western journals, also contributed to this.

This localization resulted in the further growth of national journals. In the absence of any incentive to publish abroad, plus linguistic and methodological differences deter the Armenian scientists from publications in foreign journals. Thus, a huge percentage of the science made in Armenia was closed for the world. From 1991-2016 the number of publications in WOS from Armenia was nearly 17,000, which is only 10% of total publications.

This situation raised the necessity to create the national science citation index to assess the situation in the field of science and assist integration processes.

Armenian Science Citation Index (ASCI)

The Center for Scientific Information Analysis and Monitoring was established in 2010 with an aim of unification and integration of the Armenian Science into the World Science.

One of the main tasks of the Center was the elaboration of the Armenian Science Citation Index (ASCI) (Sargsyan et all, 2010).

The goals of ASCI are the following:

- To collect the publications of the Armenian scientists in one database;
- To collect all Armenian scientific journals in one database;
- To collect the publications of the Armenian scientists referred in the international databases;
- To select best journals based on bibliometric indicators.

Nowadays there are nearly 150 scholarly journals published in Armenia. They are mainly in Armenian with or without an English language abstract. Only 3 journals – Astrophysics, Journal of Contemporary Physics (Armenian Academy of Sciences), Journal of Contemporary Mathematical
Analysis (Armenian Academy of Sciences) are indexed in the Web of Science database. Despite some works towards the increase of the Armenian journals in the WOS, this number is not yet changed. The number of journals indexed in ASCI is nearly 50 with the time frame 2007-2016. The difficulties involved at this stage are the small staff involved in the work and that a huge part of the work is being done manually. There is a problem with finding the publications to be indexed; frequently their format doesn’t correspond to the portal requirements. 

Apart from working with the databases, the Center is also involved in dissemination issues, as the idea is not easily accepted by the Armenian scientific community and needs to be spread and explained. This work is being done on daily basis on state (state bodies), institutional (scientific organizations, institutions, publishing houses) and personal levels (researchers, editors, etc.).

As a result of developing ASCI and work with different stakeholders, we have noticed a considerable raise in the Impact Factors of the Armenian journals and their overall technical and quality indicators (investment of the peer-review system). Every year the number of journals considered in the ArmJIF (Armenian Journals’ Impact Factor) is also raised (Sargsyan et al 2013).

At the same time the Center is elaborating ASCI portal with its necessary tools and supporting databases having a prototype WOS portal. ASCI portal is a multi functional information system which analysis a complete bibliometric data, abstracts and citations from the Armenian scientific journals. Such known indicators as the number of publications, number of citations, their averages, Hirsh index, and Impact factor of the journals are calculated together with dozens of other metrics. ASCI portal will comprise an Armenian Journal Impact Factor (ArmJIF) and Armenian Data Analytical Tools (ArmDAT) to make necessary comparisons, analytics, etc.

Acknowledgments
This work was done by the support of the Armenian National Fund for Education and Research (ANSEF based in New York, USA) grant № soc 4443.

References
Sargsyan Sh., Ghazaryan N., Iskandaryan E., Hovhannisyan L. & Ayyazyan N. (2010). Elaboration of a system for statistical analysis of science in Armenia according to the Armenian Science Citation Index. Bulletin of the Center for Scientific Information Analysis and Monitoring 1, 53.
Analyzing the citation impact of scholarly books based on BKCI

Liu Xiaojuan 1  Ma Liang 2  Yu Mengxia 3

1 lxi_2007@bnu.edu.cn Beijing Normal University, School of Government, Xinjiekouwai Street NO.19, Haidian District, Beijing (China)

2 bnumliang@mail.bnu.edu.cn Beijing Normal University, School of Government, Xinjiekouwai Street NO.19, Haidian District, Beijing (China)

3 bnuymx@163.com Beijing Normal University, School of Government, Xinjiekouwai Street NO.19, Haidian District, Beijing (China)

Introduction
This study analyses the citation evolution of scholarly books based on the Book Citation Index (BKCI) from two perspectives: year of publication and discipline. We have selected all the scholarly books, which were collected by BKCI, published from 2005 to 2015. This paper is organised in this manner. In first section, we describe the annual variation about the publication quantity and citations of scholarly books in different disciplines. Second section is dedicated to analyse the citation frequency distribution. The last section presents the citation curve of scholarly books from the aspect of literature age.

Annual variation about the publication quantity and citations of scholarly books in different disciplines
The quantity of scholarly books shows an increasing trend in the early years and relatively stable in recent years and varies from disciplines. Scholarly books in Social Sciences, Arts & Humanities have been published in large quantities, but their citations are much lower. In contrast, in Life Sciences & Biomedicine, the citations of scholarly books are significantly higher than other disciplines, while the publication quantities are low.

Figure 1. Annual variation about the Publication Quantity in different disciplines

Figure 2 shows that the scholarly books which published in early years have more citations than the recent ones, and it also indicates the degree of influence from publishing year varies from discipline to discipline.

Figure 2. Annual variation about the citations in different discipline

Figure 3 reflects the zero citation rate of scholarly books. The diagram shows it presents an increasing trend. In early years, the zero citation rate rises slowly year by year, but in recent years, it increases rapidly. There is little difference between the disciplines although the proportion in Social Science and Art and Humanities are higher than other disciplines.

Figure 3. Annual variation about the uncited rate in different disciplines

Citation frequency distributions
The citation frequency distributions of scholarly books have shown the strong polarisation. The share of the lowest and the highest cited frequencies is over 60% and nearly 10%, which is much higher than the other citations intervals.
Across different disciplines, scholarly books in Life Sciences & Biomedicine and Physical Sciences performed better than other disciplines in the highest cited frequencies interval, while scholarly books which have the lowest cited frequencies are more likely in Social Sciences and Arts & Humanities.

Publication year is another important factor. Data in Table 1 is calculated using the formula below:

\[
D = R_I(y) - R_J(y) \quad \text{for } i \neq j, \quad y \in [0, 10] + \infty
\]

In the formula above, \( y \) means the publication year, \( R_I(y) \) means the citation frequency of the books published in year \( y \) received in year \( j \), \( R_I \) means the proportion of books which citation is \( bij \) since published.

We found that the earlier the scholarly books published the more citations it will receive, especially in Physical Sciences and Life Sciences & Biomedicine. The polarisation of citation frequency distributions is show much stronger in the earlier publication years in Social Sciences and Arts & Humanities.

Table 1. Difference between the share of scholarly books with low and high citations

<table>
<thead>
<tr>
<th>Difference (%)</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Science</td>
<td>-37.5</td>
<td>-22.1</td>
<td>-21.0</td>
<td>-16.0</td>
<td>7.2</td>
<td>18.4</td>
<td>23.2</td>
<td>33.9</td>
<td>55.4</td>
<td>77.4</td>
</tr>
<tr>
<td>Life Sciences &amp; Biomedicine</td>
<td>-28.8</td>
<td>-22.7</td>
<td>-11.9</td>
<td>-5.2</td>
<td>14.7</td>
<td>20.8</td>
<td>21.1</td>
<td>30.0</td>
<td>49.5</td>
<td>67.0</td>
</tr>
<tr>
<td>Technology</td>
<td>-11.4</td>
<td>2.7</td>
<td>0.6</td>
<td>5.0</td>
<td>26.0</td>
<td>32.9</td>
<td>36.7</td>
<td>48.5</td>
<td>67.9</td>
<td>83.8</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>13.4</td>
<td>13.8</td>
<td>20.1</td>
<td>30.5</td>
<td>39.7</td>
<td>51.3</td>
<td>57.3</td>
<td>68.3</td>
<td>82.3</td>
<td>94.3</td>
</tr>
<tr>
<td>Arts &amp; Humanities</td>
<td>33.5</td>
<td>31.8</td>
<td>35.0</td>
<td>51.3</td>
<td>58.7</td>
<td>67.6</td>
<td>72.2</td>
<td>80.0</td>
<td>90.0</td>
<td>97.0</td>
</tr>
</tbody>
</table>

Life-time curve of citations

Uncited Rate

Scholarly books in different disciplines reveal similar uncited rate characteristics. Most scholarly books would receive at least one citation in four years.

Citation Curves

Figure 6 indicates that the citations obtained in a given year varies significantly with the age among all disciplines.

Life Science & Biomedicine, and Physical Sciences, the citation frequencies increased rapidly since the book published, and when reach to the peak, the annual citations drop slowly year by year. However, the peak appears in different years between the two disciplines. The other three disciplines do not have obvious peak in the statistical period, the citations increase with the age.
Topic modelling based network maps in cardiovascular research

Diane Gal¹, Bart Thijs², Karin Sipido³ and Wolfgang Glänzel⁴

¹ diane.gal@kuleuven.be
Department of Cardiovascular Sciences, KULeuven, Leuven (Belgium)

² bart.thijs@kuleuven.be
KU Leuven, FEB, ECOOM, Leuven (Belgium)

³ karin.sipido@kuleuven.be
Department of Cardiovascular Sciences, KU Leuven, Leuven (Belgium)

⁴ wolfgang.glanzel@kuleuven.be
ECOOM and Dept. MSI, KU Leuven, Leuven (Belgium)
Dept. Science Policy & Scientometrics, Library of the Hungarian Academy of Sciences, Budapest (Hungary)

Introduction

Structural mapping and the detection of topics together with their dynamics within a given field of research remains a challenge within scientometrics. Journal-level classification systems do miss the required granularity and fail to adequately describe specific areas within research and so article-level information must be utilised, including lexical and citation data.

Previous studies have shown that using a hybrid approach, combining information from both lexical and citation data, produces more relevant findings at this level of analysis (Boyack et al., 2011; Boyack & Klavans, 2010; Janssens, 2007; Janssens, Glänzel, & Moor, 2008; Liu et al., 2010; Liu, Glänzel, & De Moor, 2012; Šubelj, van Eck, & Waltman, 2016). Thijs et al (2017) showed that combining similarities based on noun phrases extracted from titles and abstracts using Natural Language Processing (NLP), with bibliographic coupling similarities provided further insight into the topics studied and changes over time. Boyack et al (Boyack et al., 2011), when mapping two million publications from PubMed, also concluded that a hybrid technique using NLP and BC resulted in more accurate findings.

However, discrepancies between the lexical and citation based links remain present in the current approaches. The density and thus the underlying degree distributions deviate from each other. Due to quite common lexical expressions it is more likely that papers are connected than through the references to highly cited papers. Term frequency weightings like TF-IDF solve this problem only after the calculation of the similarity and after the application of a chosen threshold. Extending the list of stop words for removal might be suggested but the list of these common expressions can be long and contain meaningful phrases, removing them would in many cases destroy strong document links. In this paper, we propose an alternative approach for feature selection as an input to similarity calculations. We will use topic modelling in a large dataset as a mean to provide the necessary data to create lexical documents networks to solve the issue of higher density. The objective of this paper is to present the method and preliminary results obtained from applying Latent Dirichlet Allocation (LDA) to create a topic model and topic network map based on a large cardiovascular dataset.

Methods

Dataset and source

We obtained the publication meta-data in the cardiovascular field (2004-2013) from the Web of Science database, based on an in house license held between ECOOM and Clarivate Analytics. Of the 478,436 articles, letters, reviews and notes in the cardiovascular field that were identified, as previously described (Gal, Glänzel, & Sipido, 2016), 478,006 contained abstracts. We pre-processed the abstracts to extract noun phrases and term shingles using Stanford NLP (see Thijs et al 2015 for more details). We removed all copyright notices at the end of the abstract, all expressions containing numbers and all white spaces to reduce the overall list to almost four million unique phrases. We then removed the most commonly used phrases based on the h-core (those expressions -in a descendant ordered list- which rank is lower than their number of occurrences).

Topic modelling

Next, we applied Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003) to the document-term matrix using MalletR to obtain the beta and gamma for the topic-term and topic-document probabilities, respectively.
**Topic-document clustering**

The topic-document matrix containing the probabilities that a document is related to a particular topic is a very dense matrix but holding often very small probabilities in the cells. Straightforward salton cosine calculations would result thus in an almost complete network. Moreover, documents that have very small probabilities on all topics are very likely to form a densely connected clique with very high similarities. Those documents have to be removed from the set. In fact, we replace all probabilities below the topic average in the matrix by a missing value. Several documents will have such missing values on all topics and will become singletons in the document network instead of ending up in a strong cluster.

Next we use Locality Sensitive Hashing (Ravichandran et al 2005) for the calculation of the cosine similarity. The main advantage of LSH in this context is that through the use of randomization functions the actual calculation is only restricted to those pairs of documents that are very likely to be similar. Through this restriction we do not calculate cosines between pairs of documents with very low similarity as these links would be removed afterwards from the networks as they have a strength below a given threshold.

**Findings**

**Topic modelling**

In our preliminary review, the top groups of terms identified for each of the 200 topics, seemed to represent coherent topics in cardiovascular research or in methods applied in cardiovascular research. A limitation of topic modelling, however, is the need to define the number of topics to set apriori, requiring qualitative review to assess the validity, overlap and granularity of the topics.

**Table 1. Topic Model Overview**

<table>
<thead>
<tr>
<th>LDA</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of publications</td>
<td>476,154</td>
</tr>
<tr>
<td>Number of terms</td>
<td>3,946,268</td>
</tr>
<tr>
<td>Number of topics*</td>
<td>200</td>
</tr>
<tr>
<td>Average number of terms</td>
<td>73</td>
</tr>
<tr>
<td>with topic with beta&gt;0.001</td>
<td></td>
</tr>
<tr>
<td>Average number of topics</td>
<td>6</td>
</tr>
<tr>
<td>with gamma&gt;0.01</td>
<td></td>
</tr>
</tbody>
</table>

*the number of topics is set apriori by the researchers

**Conclusions and next steps**

Next we will use community detection techniques like the Louvain clustering method and infomap to detect topic clusters in the documents. It is expected that the application of LDA for feature extraction will produce a network with properties more in line with bibliographic coupling based networks than using other lexical feature extraction methods. And that the topic detection is less sensitive to problems like resolution limit and degeneracy.

**References**


Science and Technology Indicators of Microalgae-Based Biofuel Research.

Pierre-Louis Gorry¹, Marcia Morales Ibarria², Philippe Gorry³

¹pilou.gorry@gmail.com, ²mmorales@correo.cua.uam.mx
Departamento de Procesos y Tecnología, Universidad Autónoma Metropolitana Cuajimalpa 05348, Av. Vasco de Quiroga 4871, Sta Fé, Ciudad de México, CDMX, (Mexico).
²philippe.gorry@u-bordeaux.fr
GREThA UMR CNRS 5113, University of Bordeaux, Ave Léon Duguit, 3608 Pessac (France).

Introduction
Microalgae are considered one of the most promising sources for biofuels (Wijffels & Barbosa, 2010). However, this alternative to fossil fuel must overcome technological and economical barriers. To gain insights into the evolution of research on biofuel production from biomass, scientometrics characterization of biomass (Chen & Ho, 2015), waste recycling research (Garechana et al., 2015) as well algae and biodiesel patenting trends (Preiss & Kowalski, 2010) have been reported. These studies have highlighted the exponential growth of bioenergy research and the dominant position of the USA in this area. However, in many countries the invention of biofuel technologies slowed down by year 2008 (Albers et al., 2016). A first attempt to link the evolution of biofuel science with R-D expenditure has been published (Azadi et al., 2017) but, none of these previous studies embraced at the same time science, technology and economic indicators to describe the evolution of biofuel innovation.

The purpose of this work is to describe biofuel research and technology trends focusing on microalgae-based biosource through a multidimensional analysis to reveal influential factors (scientific, technologic, economic or environmental).

Materials & Methods
Literature search on microalgae and biofuel research was conducted in Scopus® database until 31/12/2015, with a query across the title, abstract or keyword fields in all types of records using a string of keywords associating biofuel and other synonyms (n=6), and the major species of microalgae (n=65). Patents with priority date before 31/12/201 were retrieved from the INPADOC family patents worldwide collection using Questel® software with the same search query used in Scopus®. Patent bibliographic data were analysed with the Intellixir® software, and metrics based on patent family. Press releases were searched within the Factiva® database across the title and abstract using a simplified version of the search query run in Scopus®.

Macroeconomic, and green indicators were gathered from the OECD statistics web portal: Gross Domestic Product (GDP, in US$/capita), biofuels patent (family number), environmental public investment (US$ Millions, PPP 2010). Other indicators, such as data in relation to the oil production was gathered from U.S. Energy Information Administration (https://www.eia.gov/), or the sunshine duration (hours by year for capital city) were retrieved for each country from U.S. National Oceanic & Atmospheric Administration (http://www.noaa.gov). Descriptive statistics were run for the different indicators and some inferential statistics were performed with the help of XLSTAT® software in order to describe association or correlation between science, technology, and other indicators.

Results

Science indicators
A corpus of 163,006 and 302,774 documents were respectively collected for microalgae and biofuel research fields. The intersection of these two sets gives a final corpus of 25,712 documents defining microalgae-based biofuel research field. This research domain took off in the ‘70’ followed by an exponential growth in 2010 (Figure 1; black line).

Figure 1. Microalgae biofuel innovation trends.

The contribution of different countries ranks US first, without any dominant position, along with
BRICS countries in the top 10 (Table1). The main publishing research institution is the Chinese Academy of science (data not show).

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Paper number</th>
<th>Paper %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>US</td>
<td>5713</td>
<td>22,22</td>
</tr>
<tr>
<td>2</td>
<td>CN</td>
<td>2424</td>
<td>9,43</td>
</tr>
<tr>
<td>3</td>
<td>DE</td>
<td>1613</td>
<td>6,27</td>
</tr>
<tr>
<td>4</td>
<td>FR</td>
<td>1531</td>
<td>5,95</td>
</tr>
<tr>
<td>5</td>
<td>UK</td>
<td>1473</td>
<td>5,73</td>
</tr>
<tr>
<td>6</td>
<td>CA</td>
<td>1459</td>
<td>5,67</td>
</tr>
<tr>
<td>7</td>
<td>SP</td>
<td>1435</td>
<td>5,58</td>
</tr>
<tr>
<td>8</td>
<td>IN</td>
<td>1133</td>
<td>4,41</td>
</tr>
<tr>
<td>9</td>
<td>AU</td>
<td>1029</td>
<td>4,00</td>
</tr>
<tr>
<td>10</td>
<td>JP</td>
<td>915</td>
<td>3,56</td>
</tr>
</tbody>
</table>

**Table 1. The most publishing countries.**

**Technology indicators**

A collection of 2,992 patent families was retrieved. Microalgae-based biofuel innovation trend was measured based on patent filing priority date (Figure 1; dashed line). Then countries were ranked according to their patent family number: surprisingly, China came first, followed by US and Korea (data not show). In order to model the technology transfer efficiency, a linear regression was run to explore the relationship between the patent family number (Y) and publications number (X) as explanatory variable (Figure 2). While we can distinguish between high and low performing countries, the coefficient of correlation remains low ($r^2=0,542$) leaving us the existence of other independent variables.

**Macroeconomic & Environmental indicators**

Multiple regressions were run with various indicators (scientific, technologic, economic or environmental). An example of such analysis run on a sample of 10 countries (including DE, DK, FI, FR, IL, IT, KR, MX, PT, UK) is presented (Table 2): there is a strong correlation between the microalgae biofuel innovation (Biogas) and the overall patenting of a country (patent), sun exposure is negatively correlated with GDP/capita and research output (publications), and there is a moderate but significant correlation between green investment and oil production.

**Discussion**

Our results indicate that behind the steady increase of microalga biofuel publications over the last decade, there is some discrepancy with patent filing which could be linked to economic or policies issues. One might think that countries with important sun exposure in favour of microalga photosynthesis, would be research intensive while petroleum exporting countries would not be interested in. Our preliminary results show a more complex relationship between scientific and technological factors, and macroeconomic or environmental factors. This should be explored further through econometric modelling. Such analysis would be useful for both researchers and policymakers to understand research trends and green innovation policy incentives.

**Acknowledgments**

P-L. G. is supported by a PhD fellowship from CONACyT (Grant 7388230) Universidad Autónoma Metropolitana Cuajimalpa.

**References**


The Dissemination of the Concept of “Interventional Radiology” in Medical and Scientific Literature

Léo Mignot¹ and Philippe Gorry²

¹ leomignot@orange.fr
Centre Emile Durkheim UMR CNRS 5116, University of Bordeaux, 3 ter place de la Victoire 33076 Bordeaux (France)

² philippe.gorry@u-bordeaux.fr
GREThA UMR CNRS 5113, University of Bordeaux, Av. Leon Duguit, 33608, Pessac (France)

Introduction
Interventional radiology (IR) was born in the 1960s, at the interface of radiology and cardiology but its institutional recognition was late and incomplete. IR is still the subject of turf wars between different medical disciplines and it is not clear if this field will one day become an independent medical specialty.

Purpose
The aim of this study is to analyze the diffusion of the concept of "interventional radiology" using scientometric approach. The analysis of the history of the concept and its trends allows us to better understand how this medical field has attempted to institutionalize and to gain its autonomy.

Methods
A literature search on « interventional radiology » concept diffusion was conducted in Scopus and Factiva databases for medical science publications, and press releases respectively. Then, we analysed the resulting corpus (n=26655 documents) by conducting a bibliometric, network and descriptive statistics analysis with various analysis tools (Intellixir, Excel). Thus, we were able to examine the history of the concept and to identify the actors involved in the field (authors, journals, institutions and countries) in order to understand the institutionalization process of this medical field.

Results
The first article using the term “interventional radiology” was published by S. Wallace in 1976.

The distribution of keywords associated with the IR concept in publications reveals the predominance of endovascular techniques (angiography, angioplasty, catheters) and interventional cardiology, or “Medical Imaging” and cancer issues. Nevertheless, the temporal distribution of the publications suggests that the concept was rarely used in the literature until the 1990’s. Therefore, the founding fathers of the specialty are not the ones whom have popularized the use of the label “interventional radiology” in the literature (data not show). The growth of publications accelerated in the 2000’s, with a peak of public interest in the general press (Figure 1). The concept is mostly used in American (32%), European (34%) and Japanese (6%) articles. The use of the term in Chinese publications progresses from the late 2000s (Figure 2).

The top publishing institutions are mostly American and British ones (Table 1): the notable exceptions are the University of Seoul (2nd rank), Tokyo University (6th rank) and Toronto University (10th rank).

<table>
<thead>
<tr>
<th>Rank</th>
<th>Institutions</th>
<th>Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>New York Univ.</td>
<td>472</td>
</tr>
<tr>
<td>2</td>
<td>Univ. of Seoul</td>
<td>467</td>
</tr>
<tr>
<td>3</td>
<td>Cambridge Univ.</td>
<td>345</td>
</tr>
<tr>
<td>4</td>
<td>Stanford Univ.</td>
<td>248</td>
</tr>
<tr>
<td>5</td>
<td>Univ. of Pennsylvania</td>
<td>234</td>
</tr>
<tr>
<td>6</td>
<td>Tokyo Univ.</td>
<td>227</td>
</tr>
<tr>
<td>7</td>
<td>Univ. College of London</td>
<td>221</td>
</tr>
<tr>
<td>8</td>
<td>University of Chicago</td>
<td>212</td>
</tr>
<tr>
<td>9</td>
<td>Univ. of Los Angeles</td>
<td>202</td>
</tr>
<tr>
<td>10</td>
<td>Toronto Univ.</td>
<td>200</td>
</tr>
</tbody>
</table>

Figure 1. IR publications and press trends.

Figure 2. IR publications by country over time
Various institutions are now entitled “interventional radiology” mainly in the USA, in Europe (DE, IT, UK, FR) and in Asia (CN, JP) (data not show).

Table 2. Main journal publishing IR articles.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Journal</th>
<th>Medical field</th>
<th>Pub.</th>
<th>IF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>J Vasc Interv Radiol</td>
<td>IR</td>
<td>2052</td>
<td>2.20</td>
</tr>
<tr>
<td>2</td>
<td>Cardiovasc Inter Rad</td>
<td>IR</td>
<td>1325</td>
<td>1.25</td>
</tr>
<tr>
<td>3</td>
<td>Radiology</td>
<td>Radiol.</td>
<td>1122</td>
<td>5.56</td>
</tr>
<tr>
<td>4</td>
<td>Am J Roentgenol</td>
<td>Radiol.</td>
<td>1089</td>
<td>2.47</td>
</tr>
<tr>
<td>5</td>
<td>Catheter Cardio Interv</td>
<td>Cardio.</td>
<td>996</td>
<td>1.75</td>
</tr>
<tr>
<td>6</td>
<td>Eur Radiol</td>
<td>Cardio.</td>
<td>740</td>
<td>3.40</td>
</tr>
<tr>
<td>7</td>
<td>Am J Cardiol</td>
<td>Cardio.</td>
<td>544</td>
<td>3.60</td>
</tr>
<tr>
<td>8</td>
<td>Eur J Radiol</td>
<td>Radiol.</td>
<td>527</td>
<td>1.91</td>
</tr>
<tr>
<td>9</td>
<td>Am J Neuroradiol</td>
<td>Radiol.</td>
<td>445</td>
<td>2.33</td>
</tr>
<tr>
<td>10</td>
<td>J Am Coll Cardiol</td>
<td>Cardio.</td>
<td>417</td>
<td>11.05</td>
</tr>
</tbody>
</table>

The two mains journals of the fields are the JVIR and the CVIR, but they account less than 10% of the total IR publications (Table 2).

Discussion

Analysis of publications reveals that the emergence of IR concept in the scientific literature took place several years after the birth of the specialty. Indeed, IR is an example of delayed recognition invention: it is born with the invention of angioplasty by Dotter in 1964, and the diffusion of the concept was contemporary of the awaking of his landmark paper on angioplasty (Gorry, 2016). Thus, the concept was not promoted by the founding fathers, but by new entrants. For example, C. Dotter – who is commonly known as “the father of intervention” (Payne, 2001) did not promote the concept of IR in the literature until 1981. Moreover, if Wallace (1976) is the first to use the term IR, he rarely used it during his long and productive career (10 articles among 382). We can assume that the structuring of the field by scientific societies helped to build a research community and to develop a sense of belonging to the IR field.

Table 3. Foundation of IR societies

<table>
<thead>
<tr>
<th>Year of foundation</th>
<th>IR society</th>
</tr>
</thead>
<tbody>
<tr>
<td>1975</td>
<td>Society of Interventional Radiology (US)</td>
</tr>
<tr>
<td>1982</td>
<td>Japanese Society of Interventional</td>
</tr>
<tr>
<td>1985</td>
<td>Cardiovascular &amp; Interventional Radiology</td>
</tr>
<tr>
<td>1988</td>
<td>Society of Interventional Europe</td>
</tr>
<tr>
<td>2002</td>
<td>British Society of Interventional Radiology</td>
</tr>
<tr>
<td>2008</td>
<td>Chinese Society of Interventional Radiology</td>
</tr>
<tr>
<td>2008</td>
<td>Deutsche Gesellschaft für Interventionelle</td>
</tr>
<tr>
<td></td>
<td>Radiologie und minimal-invasive</td>
</tr>
</tbody>
</table>

The main IR scientific societies have only been created 10 to 20 years after the first medical developments of the specialty (Table 3). This may also explain in part why the emergence of the concept in the literature was so late.

The specialty tried to gain its scientific autonomy, founding its own journals in 1978 (CVIR) and 1990 (JVIR). But, it is a partial failure: these 2 journals account for less than 10% of IR publications.

The appearance of department entitled “interventional radiology" can be seen as another attempt of empowerment and institutionalization. This process can be interpreted as an attempt to legitimize interventional radiology. It is also an identity claim sustained by the proclamation of a clear membership in a medical specialty. More generally, the publications mobilizing the concept of IR focus mainly on endovascular procedures, medical imaging and cardiology. Those overlaps could explain why most of the IR articles are published in the fields of radiology and cardiology. It could also be explained by the desire to communicate the results in journals of those disciplines and to ignore the disciplinary boundaries.

Today, IR still remains a heterogeneous field with unstable borders. The specialty has not reached a full scientific autonomy and its institutionalization process is still ongoing. This partial failure of institutionalization of the IR field could partly be explained by the initial hybrid position of the specialty, which emerges at the crossroads of different disciplines (radiology, cardiology, surgery) (Becker, 2001).

Through the case of IR, this study contributes to our understanding of the nature and history of interdisciplinarity and how it works. It must lead us to be careful to scientometrics indicators and analysis.

Acknowledgments

This work was supported by grants from the French National Cancer Institute (INCa #6165), and the IdEx Bordeaux.

References


Introduction

Journal impact factor has become an important standard to judge the quality of scientific publications over the years, and it also influences the other evaluation system. However, it also subjects to some criticisms. More and more journals adopt the form of publishing online in advance so that readers can browse the latest scientific research in the first time, while more and more researchers will also publish their research results in some open access database, such as some social networking sites. The online publication of the journal has a lot of ways, such as the publication would like to publish some papers or a brief description of the papers on their website in the first time, these papers often are high-quality or current hot topic. These measures can attract the reader's attention to the journal content; Another form is the journal submitting the papers to the database before the date of publication in print form, then readers can retrieve the papers in databases. Nature, Science, Springer, and Elsevier have already used the "AOP", "Express", "online first", and "Inpress" date information, which are generally earlier than the publication date in print form. We only research the online publication ahead of print, which submitted to databases earlier than the date in print. We want to discuss the relationship between online advance publication and Journal Impact Factor.

Data and Methods

In this research, we estimated 284 journals’ Impact Factors during 2012-2015, and we also estimate journals’ online-to-print lags and Lag-corrected Impact Factor. We refer to the method proposed by Tort A B. While Tort A B’s paper only select 61 journals under the “Neurosciences” category, and limit a lot of conditions. So the conclusion cannot be extensive convincing. But we have richer and more general data from every category to verification the influence of advance online publication on Journal Impact Factor.

Journal Selection

We first considered all 3375 journals which we retrieved in Springer database (https://link.springer.com/search?facet-content-type=%22Journal%22). We only reserved 184 journals and excluded other journals from this list because of (a)we only analyzed journals that publication is not zero during 2010-2016,(b)we only analyzed journals that we can retrieval Journal Impact Factor and citation frequency in the Journal Citation Report (JCR) database published by ISI (http://webofknowledge.com/JCR).

Estimation of Online-to-print Lags for Selected Journals

Online-to-print lag was computed as the difference between the online appearance date of an article and its official date of publication, as obtained from Springer records. Mean online-to-print lags for journals were obtained from all published articles in a given calendar year.

**Figure 1. Online-to-print Lags**

Estimation of Journal Impact Factors

Take 2015 impact factor as an example, the 2015 impact factor is defined as the number of citations in 2015 to articles published in 2013 and 2014 divided by the number of articles published in 2013 and 2014.

Estimation of the Lag-corrected Impact Factor

To calculate the lag-corrected impact factor index, we also take 2015 Lag-corrected impact factor as an example. The 2015 Lag-corrected impact factor is defined as the number of citations in 2015 to articles published in 2013 and 2014 divided by the number of articles online appearance in 2013 and 2014. For example, since the average delay period for the journal is three months, papers that published between April 2013 and March 2015, actually would be retrieval in advance. So these articles are online from 2013 to 2014.
Results
We analyzed 284 journals, and the proportion of online publication ahead of print was 52.11%, 66.20%, 66.20%, 66.55% and 69.01%, so we consider that the online-to-print Lags has been on the rise in the past few years. According to statistics, it found that the 284 periodicals of the average delay in the distribution of 0 to 18 months, and delay month showed a gradual increase in the trend. The current journals online advance published has become a more extensive phenomenon.

Figure 3. Journals online-to-print Lags statistics

We first calculated 284 journals’ Lag-corrected Impact Factor. Then we calculated Lag-corrected Impact Factor gradient to study change. The delay month is negatively related to Lag-corrected Impact Factor gradient. If the delay month longer, the Lag-corrected Impact Factor gradient would be smaller, in another hand, Lag-corrected Impact Factor would be decrease.

$$\text{Lag-corrected Impact Factor gradient} = \frac{\text{Lag-correctedIF} - \text{IF}}{\text{IF}} * 100$$

Formula 1. Lag-corrected Impact Factor gradient

We found that Lag-corrected Impact Factor small or unchanged proportion of the total number is greater than 75%, that means if a journal always publishes their papers online in advance, its Impact Factor would be inflated.

Figure 4. 2015 Journals Lag-corrected Impact Factor gradient and delay month distribution.

We found that the bimonthly journals and quarterly’s proportion of the lag-corrected impact reducing is large, while the annual is basically unchanged. Because bimonthly journals and quarterly usually have short and timely publishing period, online publication ahead of print would have effect on impact factors. While annual have longer publishing period, so the effect is weaker. Comparing with Tort A B’s research, some results are similar, such as delays between online and print publication of articles increased steadily over the last years and online-to-print delays would be influence JIF. But we also have the difference. We found that JIF would not always increase, even though a journal always publishes their papers online in advance. Considering the current method of JIF calculation, which is mainly based on the date of publication in print form, despite the fact that most journals now make their articles available online before that date. Because we found that journal online-to-print Lags may affect impact factor, we suggest that calculation of JIF should consider the date of a paper’s online appearance, not only the date of the publication in print.

References
Introduction

Although there are significant advocacies in favour of Open Access (OA) mandates in the U.K., there are less evidence on how gold OA research (OA from publisher) draws altmetrics compared to green OA (OA from authors or repositories). Numerous research has examined conventional citation advantage of OA, indicating contradicting results (see annotated bibliography regularly being updated by Wagner since 2010). However, there are a few studies on advantages of research access models on altmetrics such as Mendeley readerships and Twitter uptakes (Adie, 2014; Alhoori et al., 2015). Therefore, as a follow-up study current research aims at comparing research impact across access models (Maleki, 2015) in a more OA-intensive country. The research is a large-scale evidence indicating whether and how altmetrics is influenced by various OA models, particularly journal papers accessible through Digital Open Access Journals (DOAJ), PubMed, and ArXiv.

Method

In order to address research question, 413,770 UK journal articles with Digital Object Identifier (DOI) out of 466,640 articles published between 2011 and 2014 was exported from Scopus. Scopus was used due to its broader coverage of Open Access publications. To identify papers’ subject and OA status in DOAJ, source index of Scopus was used. PubMed availability of papers was identified using DOI converter in nlm.nih.gov. In order to examine availability of ArXiv preprints, arXiv IDs was extracted by software developed to submit DOIs through arXiv API. Tweet count and Mendeley readers was collected using software program Webometric Analyst (lexiurl.wlv.ac.uk). All data was collected in June 2017.

Findings

Table 1 indicates that proportion of PubMed-indexed papers were just over three times (34%, 140,548) as much as articles published in OA journals (11% 47,376 articles), whereas the UK articles in Health Sciences were predominantly available through PubMed (60%). ArXiv, however, was used relatively low in Physical Sciences (only 11%, 16,738).

In order to compare means of speculated metrics (see Table 1) Mann-Whitney test was initially used, indicating that all access types significantly showed higher impact compared to non-accessible articles, despite mean readership impact of preprint-provided papers in ArXiv (20) which had slightly, but significantly less readers than non-ArXiv (24). There was a minimal but significant difference in mean Scopus citations to OA and non-OA papers, whereas both PubMed (24) and ArXiv (23) papers advantaged almost twice as many citations as non-PubMed (12) and non-ArXiv (14). Readerships were more numerous than citations while on average PubMed papers (41) were bookmarked almost double that of non-PubMed (21). Similarly, mean readerships of OA (34) was significantly above, but not much more numerous than non-OA (27). Mean non-ArXiv readerships (24) was above ArXiv (20), however. Finally, papers in PubMed (6), OA journals (5) and ArXiv (2) had significantly more average tweets than non-PubMed (1), non-OA (2) and non-ArXiv (1) all significant at p < 0.0001.

Table 1. Mean (median) geometric mean of natural log(n+1) for three metrics across access models

<table>
<thead>
<tr>
<th>Access models</th>
<th>Citations</th>
<th>Readerships</th>
<th>Tweeters</th>
<th>Total articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>OA&lt;sup&gt;a&lt;/sup&gt;</td>
<td>16 (9) 10**</td>
<td>34 (22) 21**</td>
<td>5 (1) 3**&lt;sup&gt;R&lt;/sup&gt;</td>
<td>47,376 (11%)</td>
</tr>
<tr>
<td>Non-OA&lt;sup&gt;a&lt;/sup&gt;</td>
<td>16 (7) 9&lt;sup&gt;R&lt;/sup&gt;</td>
<td>27 (15) 15&lt;sup&gt;R&lt;/sup&gt;</td>
<td>2 (0) 3</td>
<td>366,394 (89%)</td>
</tr>
<tr>
<td>PubMed&lt;sup&gt;a&lt;/sup&gt;</td>
<td>24 (12) 13**&lt;sup&gt;R&lt;/sup&gt;</td>
<td>41 (25) 25**&lt;sup&gt;R&lt;/sup&gt;</td>
<td>6 (1) 4**&lt;sup&gt;R&lt;/sup&gt;</td>
<td>140,548 (34%)</td>
</tr>
<tr>
<td>Non-PubMed&lt;sup&gt;a&lt;/sup&gt;</td>
<td>12 (6) 8</td>
<td>21 (12) 12</td>
<td>1 (0) 2</td>
<td>273,222 (66%)</td>
</tr>
<tr>
<td>PubMed&lt;sup&gt;b&lt;/sup&gt;</td>
<td>22 (10) 12**&lt;sup&gt;R&lt;/sup&gt;</td>
<td>33 (22) 23**&lt;sup&gt;R&lt;/sup&gt;</td>
<td>6 (1) 4**&lt;sup&gt;b&lt;/sup&gt;</td>
<td>90,319 (60%)</td>
</tr>
<tr>
<td>Non-PubMed&lt;sup&gt;b&lt;/sup&gt;</td>
<td>10 (5) 7</td>
<td>18 (13) 13</td>
<td>0 (0) 2</td>
<td>59,005 (40%)</td>
</tr>
<tr>
<td>ArXiv&lt;sup&gt;c&lt;/sup&gt;</td>
<td>23 (11) 13**&lt;sup&gt;R&lt;/sup&gt;</td>
<td>20 (11) 12</td>
<td>2 (0) 3**&lt;sup&gt;c&lt;/sup&gt;</td>
<td>16,738 (11%)</td>
</tr>
<tr>
<td>Non-ArXiv&lt;sup&gt;c&lt;/sup&gt;</td>
<td>14 (8) 9</td>
<td>24 (13) 13**&lt;sup&gt;R&lt;/sup&gt;</td>
<td>1 (0) 2</td>
<td>134,293 (89%)</td>
</tr>
<tr>
<td>Overall</td>
<td>16 (8) 9</td>
<td>28 (16) 16</td>
<td>3 (0) 3</td>
<td>413,770 (100%)</td>
</tr>
</tbody>
</table>

<sup>a</sup> UK journal articles (2011-2014) in Scopus with DOIs; <sup>b</sup> Only Health Sciences; <sup>c</sup> Only Physical Sciences; ** significantly higher mean than counterpart category at p < 0.0001 (e.g. OA vs. non-OA); <sup>R</sup> Access type with significant positive coefficient in regression model
As shown in Figure 1, there was a relatively moderate, but significant and slightly decreasing Spearman’s correlation between formal citations and readerships (C-R) of publications in all access forms, probably due to falling number of citations accrued by more recent papers. Correlations for PubMed papers (at around r = 0.6), which is above non-PubMed (changing between 0.5 and 0.6) was the strongest, and correlations in OA journal articles are also marginally above non-OA (both around 0.6). However, the reverse is seen in arXiv where correlations between citations and readerships in physical science (about r = 0.6 for non-arXiv) did not rise with arXiv preprints (around 0.5).

On the other hand, correlations between citations and tweeter counts (C-T) was the weakest in all access models ranging between 0.1 and 0.3, while only slightly fluctuating over years. Despite low correlations, the broadest difference is made by PubMed access, while arXiv and OA journals also had a slight rise of tweets in line with citations.

In contrast to correlations with formal citations, there are increasing correlations between readerships and tweeter counts (R-T) over the period given, which probably reflects the increasing users of Mendeley and Twitter engaging online with academic articles.

Discussions

In order to model the uptake of papers through three access models an ordinary least squares regression analysis was conducted after data was normalised using natural log of metrics plus 1. Findings of regression showed that PubMed accessibility was significantly positive predictive of citations (standardised beta = .143, standard error = .009), readerships (b = .210, e = .005) and tweet uptakes (b = .365, e = .010). Also, consistent with Mann-Whitney test, ArXiv was positive indicative of citation (b = .117, e = .009) and tweet uptake (b = .311, e = .010) but not readerships (b = .015, e = .005). However, OA journal articles only had a significantly positive tweet uptake (b = .106, e = .011) but significantly negative citation (b = -.019, e = .010) and Mendeley impact (b = -.017, e = .006), which could be due a distribution shift occurred in normalized data as means and medians were relatively close in two groups. All regression results are significant by 95% confidence interval at p < .001.

Advantage of metrics for PubMed articles and only tweets for OA articles was also seen in a previous research (Maleki, 2014), however, results are completely inconsistent in that current research denies readership advantage of ArXiv available research and rather favours citations and tweeter uptake for ArXiv preprints.

Conclusion

Repository (PubMed) and author provided (ArXiv) access models generally seem to be more successful than OA journal articles in terms of impact, due to more impact on formal citations and readerships. OA journal articles constantly showed a benefit of tweet uptakes but not significantly more citation or readership impact. Therefore, there are promising aspects of impact for research available at PubMed, both formally and informally. Articles of scientists of Physical Science in the U.K. which are highly read, however, seem not to be available on arXiv.

References


A scaling approach to tackle the heterogeneity of HEIs

Giuseppe Catalano, Cinzia Daraio, Martina Gregori, Henk F. Moed1 Giancarlo Ruocco2

1 catalano@di.uniroma1.it; daraio@di.uniroma1.it; martina.gregori@uniroma1.it; henk.moed@uniroma1.it

DIAG, Sapienza University of Rome (Italy)

2 giancarlo.ruocco@roma1.infn.it

Center for Life Nano Science, Fondazione Istituto Italiano di Tecnologia (IIT), and Department of Physics, Sapienza University of Rome (Italy)

Introduction

The assessment of the performance of Higher Education Institutions (HEIs) at the micro (institutional), meso (regional) and macro (country) level is an important and recurrent question in the higher education’s policy debate. The modernisation agenda for Higher Education in Europe (European Commission, 2016) identifies the relevance of creating effective governance and funding mechanisms for higher education among the five key priorities for this sector. It is underlined the importance to ensure greater flexibility and autonomy for institutions to specialise more easily, promoting better educational and research performance while fostering excellence within higher education systems.

Different models of governance (Agasisti and Catalano, 2006; Capano et al. 2015) are applied by policy makers trying to improve the systemic performance of Higher Education. However, the analysis of the performance of HE systems is far from being easy to deal with. One of the main critical issue to address properly the assessment of the performance, in a multi-level (systemic) perspective, is the consideration of the heterogeneity of the HEIs involved.

Among the heterogeneity factors of HEIs, the disciplinary specialization or subject mix is considered one of the most relevant (López-Illescas et al. 2011, Daraio et al. 2011).

Main objective of the study

Recently, Bonaccorsi et al. (2017) extend the results of Ruocco and Daraio (2013) to the evaluation of bibliometric indicators of Social Sciences and Humanities (SSH) and propose a scaling approach as a tool for indirect qualitative-quantitative comparative analysis across heterogeneous disciplines. In this paper we (i) demonstrate that the distributions of total enrolled students (ENR STUDENTS) of European HEIs in four main fields of education (ENG, MED, NAT, SSH) follow a Log-Normal:

\[ f(x) = \frac{1}{\sqrt{2\pi}x} \exp\left\{ -\frac{1}{2\sigma^2} \left[ \ln \left( \frac{x}{\xi} \right) \right]^2 \right\} \]

and the rankings (in percentile, then with a maximum value equal to 100) of ENR STUDENTS have been fitted with a cumulative Log-Normal distribution:

\[ F(x) = 50 + 50\text{erf}\left( \frac{\ln(x/\xi)}{\sqrt{2}\sigma} \right) \]

where:

\[ \text{erf}(.) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \]

These functions depend on two parameters, the scale factor, \( \xi \), which is a kind of exchange rate to compare heterogeneous categories of students (namely ENG, MED, NAT and SSH), and the parameter \( \sigma \), or shape parameter, that should be close to one if the compared categories are similar, that is, in the case of universality of their underlying distribution. See Figure 1.

Table 1 shows the results of the estimates, obtained by applying a non-linear fit algorithm (Levemberg-Marquardt). For each field of education (ENG, MED, NAT, SSH), the table shows the value of the estimated scale parameter \( \xi \), its Standard Error (S.E.), the standard deviation (\( \sigma \)) and its S.E.; and normalizations before running systemic performance assessment and comparison.

Data

The data analysed come from the European Tertiary Education Register (ETER) which is a database that provides a core set of data on 2,239 HEIs in 31 European Research Area countries for the years 2011-2014. Most ETER data can be freely downloaded from the project website (http://eter.joanneum.at/imdas-eter/).

Method and results

Following Ruocco and Daraio (2013) the distributions of the total enrolled students (including ISCED 5 to 7 students, average 2011-2013), called ENR STUDENTS hereafter, aggregated in four broad fields of education ENG (Engineering and Technology), MED (Medicine), NAT (Natural Sciences) and SSH, have been assumed as a Log-Normal:

\[ f(x) = \frac{1}{\sqrt{2\pi}x} \exp\left\{ -\frac{1}{2\sigma^2} \left[ \ln \left( \frac{x}{\xi} \right) \right]^2 \right\} \]

and the rankings (in percentile, then with a maximum value equal to 100) of ENR STUDENTS have been fitted with a cumulative Log-Normal distribution:

\[ F(x) = 50 + 50\text{erf}\left( \frac{\ln(x/\xi)}{\sqrt{2}\sigma} \right) \]

where:

\[ \text{erf}(.) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \]

finally the statistics of the reduced Chi-square and the Adj. R-square.

![Figure 1. Cumulative scaled distributions of ENR STUDENTS by field of education (ENG, MED, NAT, SSH). The figure shows the distributions (Eq. 2) of ENR STUDENTS by fields of education divided by their respective scale parameters, ξ, whose values are reported in the legend of the figure.](image)

The scale parameter ξ, acts like a kind of exchange rate to compare the heterogeneous categories of students. The standard deviation (σ) estimated are all close and around one, showing that we are in presence of universality. Finally, it is worth noting that we have a very high fit for all the considered categories, as all the Adjusted R-Squares are almost equal to one. This result confirm that ENR STUDENTS have a Log-Normal distribution across the different fields of education considered. Figure 2 illustrates the obtained fit.

![Figure 2. Fit of the cumulative distributions of ENR STUDENTS by fields of education.](image)

Discussion and conclusions

The results of this analysis show that to make appropriate comparisons it is important to consider the heterogeneity of the disciplinary mix of education. The estimated scale parameters by field of education represent sound thresholds for comparing the different categories of enrolled students among HEIs and at the country level.

The scaling approach we propose in this paper could be an interesting tool to analyse and tackle the different sources of heterogeneity of HEI systems included in multi-level models of performance where country level statistical data are combined with micro-level institutional data. It is worth then to further corroborate and apply the scaling approach proposed here to implement appropriate normalization strategies in multi-level systemic performance evaluation models.

Acknowledgments

The financial support of the Italian Ministry of Education and Research through the PRIN Project N. 2015RJARX7 is gratefully acknowledged.

References


<table>
<thead>
<tr>
<th>Field Edu.</th>
<th>ξ (S.E.)</th>
<th>σ (S.E.)</th>
<th>Reduce Chi-Sqr</th>
<th>Adj. R-Square</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENG</td>
<td>1383 (2.25)</td>
<td>1.258</td>
<td>1.491</td>
<td>0.998</td>
<td>979</td>
</tr>
<tr>
<td>MED</td>
<td>1319 (3.19)</td>
<td>1.246</td>
<td>2.674</td>
<td>0.997</td>
<td>782</td>
</tr>
<tr>
<td>NAT</td>
<td>945 (3.27)</td>
<td>1.272</td>
<td>2.898</td>
<td>0.997</td>
<td>889</td>
</tr>
<tr>
<td>SSH</td>
<td>2011 (3.93)</td>
<td>1.561</td>
<td>2.524</td>
<td>0.997</td>
<td>1763</td>
</tr>
</tbody>
</table>

Table 1 Results of the fit by Field of Education of the Enrolled students.

*Note: The red curve represents the fit to the Log-Normal cumulative distribution (Eq. 2).*
TYPE: Posters

Bibliometric Research of Faculties based on Local Databases

Veslava Osinska¹ Piotr Malak²

¹wieo@umk.pl
Nicolaus Copernicus University, Torun (Poland)

piotr.malak@uwr.edu.pl
University of Wroclaw, Wroclaw (Poland)

This research is sponsored by Polish National Science Center (NCN) under grant 2013/11/B/HS2/03048/Information Visualization methods in digital knowledge structure and dynamics study.

Introduction

The state of under-representation of the humanities and social sciences in global databases such as Web of Science and Scopus has become permanent and detrimental. But for scientometric research on mezzo (institutional) level local databases can successfully substitute world indexes. Authors present a case study of gathering and visually analysing of unstructured data deriving from NCU Website. Cleaning and processing the data allow to identify metadata which describe academic activities connected with organizing different events at University. Another data source is University bibliographic database Expertus, the most complementary information resource concerning the publications of employees. This way, two different activities: publishing and organisational one can be compared for each faculty or equivalent division and thus evaluate collective impact in academic environment.

Data and metadata

Data collecting

Application form accessible at University Website https://www.umk.pl/badania/konferencje/ allows to collect such metadata about conferences as: Date, Title, Faculty, Organiser(s), Place, Leader, Secretary, Description. Downloaded data due to HTML format required semiautomatic cleaning. The relatively small records quantity \( N = 1344 \) allowed for quick estimation of similarities and relationships. Due to number of organizing units (Organiser(s) field), an asymmetric matrix has been created for the quantitative representation of the intensity of cooperation in conference organizing. Asymmetry is caused by unequal roles of the organizer and co-organizer.

Data aggregation

All identified University units (over 300) were grouped into the main categories corresponding to 17 Faculties and one additional, General. Thus, for publishing and organisational activity, cumulative data divided into 18 categories was obtained. In order to compare the received visualizations with the generated image according to the Web of Science data, the next level of grouping faculties into scientific areas was performed. Ultimately, in our study we involved domains, such as:

- humanities,
- natural science,
- life sciences,
- medicine/clinical sciences,
- engineering and technical science,
- social sciences.

Analysis

Graphs, maps and diagrams were used in the visualizations. To illustrate the dynamics of activity in the areas of conference organization area chart is used as the most effective in the trend study (Fig.1).

Figure 1. Dynamics of faculties publishing per year.

The band width exposes publication quantity each year for each faculty. It allows to observe all changes and parallel to capture information which is the most effective unit for analyzed period. For comparison local and global databases, ring charts were used. Figure 2 presents domain structure for three data sources. As mentioned before, standardized categorization was applied. All domains colored by scheme, which is available at InCites reports: https://incites.thomsonreuters.com/. Outer ring at Fig. 2 shows the best representation of publications for humanity and social sciences according local database –Expertus.
The visualization of social activities working on joint ventures (conferences, symposia, seminars) can be accomplished by using graphs (Garfield, 1994). We used the legible kind of graph – circular network layout (Fig. 3). Band width indicates the intensity of cooperation in organizing conferences. This way we discovered the close relations in such kind of activity between faculties (for example history and political science).

In order to prove this observation NLP techniques were used. Texts of conferences description were processed by Ward’s clusters analysis (1963) and visualized - the result we can see on Figure 4. Dendrogram presents faculties similarity based on texts clustering. Visualisation results confirmed previous assumptions according relationships between organizational units.

The set of metadata predisposes for extending mapping and broadening the perspective of study.

We can also carry out geomapping and provide analysis of foreign cooperation of University. We can compare publishing versus mobility for each year, but all maps exceed the scope of current work.

Figure 2. Domain structure of publications of University employees according three databases.

Figure 3. Co-organizing local events by different faculties.

Figure 4. Co-organizing local events by different faculties.

Summary
Authors present analytical possibilities of local database in confrontation with global scientometric indexes. Humanity and social sciences became visible on science map based on local data. Visualisations techniques were chosen according datatypes and analysis purposes (Osinska & Malak, 2016). Thus, circular graph shows cooperation between units, area chart – dynamics, ring diagram – comparison and structure and dendrogram – close and far similarity. Authors proposed publishing activity of scientists to complement by alternative one - organisational and this way to broaden evaluation framework.

References
Garfield, E. (1994) Scientography: Mapping the tracks of science. Social & Behavioural Sciences 7 (45), 5-10;
Some Reflections to China’s International Collaboration

Yi Han¹, Yanxiao Liu², Lanni Shen³ and Bihui Jin⁴

¹hanyi72@swu.edu.cn, ²1208913462@qq.com, ³1837495883@qq.com
⁴jinbh@mail.las.ac.cn

¹²³College of Computer and Information Science, Southwest University (P R China)
⁴National Science Library, Chinese Academy of Sciences (P R China)

Abstract
This paper is to reflect the China’s international collaboration in science during the period 1995-2015 and to discover the dynamics and the channel of international collaboration between China and the main scientific and technological powers. USA, Japan, Germany, UK, and France are chosen the international target countries to compare, the collaborative data of Big 6 are retrieved in Web of Science, and some descriptive statistical analyses are given. The data show that China’s share of internationally co-authored papers is still considerably lower than that of the other main countries. The numbers of collaborative papers in five nations have grown much faster than the numbers of non-collaborative ones. In China collaborative papers grow only slightly faster than non-collaborative ones, especially over the latest years. China’s international collaboration in science has obvious disciplinary preferences, focusing mainly on traditional fundamental disciplines. Overseas Chinese play an extremely important role in mainland China’s international collaboration, providing an optimal channel to overcome linguistic and cultural barriers.

Keywords
International scientific and technological collaboration; China’s international collaboration; overseas Chinese phenomenon

Conference Topic
Country-level studies

Introduction
By the end of the fifties in the 20th century, Smith (1958) noted a trend towards more and more multiple authorship. This led Price (1963) to predict the demise of the single-author article. Although this prediction has not come true, the number of single-author articles has indeed steadily declined. Wuchty et al. (2007) confirmed the general increase of the influence of teams in science, including that by this phenomenon the process of knowledge creation has fundamentally changed over the years. Scientific collaboration can be described as a process by which scientists with common interests work together to create new knowledge (Katz, 1994, 1997; Coccia & Wang, 2016). It is generally accepted that team work increases quality and productivity and hence the received number of citations, although collaboration does not necessarily lead to success, such as a lot of citations. (Figg et al, 2006; Glänzel, 2008; Elango et al, 2015). Generally speaking, nowadays a scientist is not anymore an isolated actor but one member, among many, of an international research team, working together to discover the rules of nature and society (Cronin, 1982; Finardi & Buratti, 2016). Due to differences in research accumulation during investigations there
are always leading teams and followers (Zhou & Bormman, 2015; Ma et al., 2016). If researchers collaborate with the leaders, knowledge and technical barriers are reduced and this helps them to reach the research front. Yet, leaders too may benefit from collaborating, as increased manpower can greatly accelerate the development and diffusion of new knowledge, new technologies and new methods.

Developing countries can greatly extend their scientific productivity and strength, in the same time increasing the visibility of their researchers, by collaborating with leading countries (Rigby & Edler, 2005; Jonkers & Tijssen, 2008; Zhou et al., 2009; Ynalvez & Shrum, 2011; Wagner et al., 2015). Collaboration gives, moreover, direct access to complementary knowledge or skills, unique sites and facilities, and shares costs and risks (Kim, 2005; Birnholtz, 2007; Haustein et al., 2011). Starting as a developing (or at least a not-yet-fully-developed) nation China’s science and technology has increased considerably during the latest twenty years. The exponential increase in the absolute and relative number of WoS papers bears testimony of this fact (Jin & Rousseau, 2005). Yet, outstanding scientific and technological discoveries are not distributed evenly over countries, regions or provinces (or states) within a country (Wang et al., 2005; Zhou et al., 2009). By studying international collaboration we intend to obtain a better insight in the position of China with respect to other big countries.

We would like to add one more point, namely that collaborative research may not only lead to publications but also to patents (Chen et al., 2013). This aspect is not the topic of this paper. Hence, the main purpose of this article is to describe China’s position in the international fray, restricted to its output in scientific papers.

**Data and Methods**

The articles jointly written by several authors reflect scientific and technological collaboration. Consequently international collaboration is reflected by articles jointly published by authors in different countries. Articles co-authored by Chinese scientists in the context of international collaboration are mainly published in international journals. Zheng et al. (2012) mentioned that in domestic journals only 1.5% of the publications are the result of international collaboration. Hence, if one wants to study China’s international collaboration as shown by published articles one must make use of an international database. Consequently we use Thomson Reuter’s Web of Science (WoS). Publications are restricted to the article and review type.

The United States, the UK (taking care of the fact that the UK consists of England, Scotland, Wales and Northern Ireland), Germany, Japan and France are, together with China the countries producing the most articles. They will be referred to as the Big 6. In 2015, the total articles’ number recorded in WoS is 2,632,093, and the number produced by the Big 6 is 1,529,243. Nowadays we focus on these six countries. Taking the names of the chosen countries as retrieval terms, the retrieved data are divided into two groups: international collaborative papers on the one hand and non-international collaboration on the other (authors are all from China, including single author papers).
By choosing the above mentioned countries we expect to grasp the main trends and characteristics of international collaboration with special emphasis on the role played by China. Furthermore, we explore the dynamics of international collaboration and try to find the optimal path for China’s scientists when it comes to participation in international scientific and technological endeavors. Occasionally we will refer to papers written as the result of an international collaboration as “co-papers” (as a short-hand).

Because of this aim, the WoS data set of articles and reviews in 1995-2015 are retrieved. Besides all papers (of article and review type), we also consider separately the collaborative papers, written in collaboration with the main countries with which these countries collaborate the most. Articles by overseas Chinese are especially collected as well as the fields in which collaboration occurs. Data were collected in September 2016.

**Results**

*International collaboration trends of the BIG 6 countries*

Taking the name of Big 6 as retrieval terms, the retrieved data are divided into two groups: international collaborative papers, of which the affiliation consists of two or more than two countries, and non-international collaborative one, of which the affiliation only one country. The full counting is recorded every one year since 1995, and the growth rate is calculated respectively. The calculation results are shown in Table 1.

<table>
<thead>
<tr>
<th>Year</th>
<th>Non-collaborative papers</th>
<th>Co-papers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>USA</td>
<td>CHINA</td>
</tr>
<tr>
<td>1997</td>
<td>-3%</td>
<td>26%</td>
</tr>
<tr>
<td>1999</td>
<td>25%</td>
<td>48%</td>
</tr>
<tr>
<td>2001</td>
<td>0%</td>
<td>51%</td>
</tr>
<tr>
<td>2003</td>
<td>26%</td>
<td>59%</td>
</tr>
<tr>
<td>2005</td>
<td>-4%</td>
<td>46%</td>
</tr>
<tr>
<td>2007</td>
<td>3%</td>
<td>28%</td>
</tr>
<tr>
<td>2009</td>
<td>8%</td>
<td>26%</td>
</tr>
<tr>
<td>2011</td>
<td>3%</td>
<td>22%</td>
</tr>
<tr>
<td>2013</td>
<td>4%</td>
<td>32%</td>
</tr>
<tr>
<td>2015</td>
<td>-3%</td>
<td>24%</td>
</tr>
</tbody>
</table>
The data in Table 1 show that, except for China, the growth rate of international collaborative papers is much faster than the growth rate of non-collaborative ones. What could be the reason for this phenomenon?

When a society matures, the division of labor and in particular the numbers of practitioners engaged in scientific and technological research may stay constant, leading to a stabilization of its scientific and technological output. In order to maintain their international position developed countries must adapt their scientific policies and try to increase their attractiveness for foreign scientists, heavyweights as well as average ones. These and other factors must be optimized to maximize scientific and technological output. Supported by a national R&D policies, international collaboration promotes scientific development and progress, leading to prosperity in the collaborating nations. For these reasons developed states have vigorously strengthened international collaboration over the past twenty years. Promoting scientific and technological collaboration has become an important point in the research policies of most countries all over the world.

In contrast, China is not yet a fully developed country, and its society is changing at full speed. On the one hand, the number of practitioners engaged in scientific and technological activities in China continues to increase, and the number of scientific papers maintains a sustained high growth. On the other hand, there is no doubt that it is a good choice to promote scientific and technological development through international collaboration. Therefore, all Chinese papers (internationally collaborated or not) maintained a double-digit growth during the recent 20 years. Furthermore, while China’s growth rate of non-collaborative papers tends to decrease after the peak in 2003; its international collaborative papers kept a sustained strong increase, and its growth rate is also higher than that of the other five nations. But until 2015, China’s scientific output(330,927) is still only about half of that of USA(646,528). As a not-yet-fully-developed country, it is of great importance for China to develop a sound scientific policies and to promote scientific research, including international collaboration, to follow or even lead world trends.

The role played by China in international collaboration

In international collaboration, the ranks of authors in the byline of papers usually reflect the contribution of participants, reflecting role differences in collaboration. In particular, the corresponding author is often the person responsible for the research project. Therefore, by analyzing whether an author in collaborative papers is the corresponding author or not, the scientific contribution of authors in such papers can be analyzed. In this section, we investigate if China undertook either a leading role or
Table 2. Proportion of corresponding authors (CAs) in international collaborative papers in 2007 and 2011

<table>
<thead>
<tr>
<th>nation</th>
<th>2007</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Co-paper</td>
<td>CAs</td>
</tr>
<tr>
<td>China</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>20324</td>
<td>10975</td>
</tr>
</tbody>
</table>

Table 2 shows that the proportion of corresponding authors in China’s international collaboration is over 50%, and staying relatively stable in 2007 and 2011. Of course, the simply descriptive statistical data only give us some superficial impression, and the in-depth analysis need to use panel analysis tools to reveal the rules.

Which role did China take in the collaboration with the five other important nations? Table 3 shows that in collaboration with major scientific and technological powers, the proportion of corresponding authors is slightly lower than the average of China in 2007 and 2011, especially in collaboration with Germany and France. Therefore, in order to enhance China’s capacities in international scientific and technological collaboration, the proportion of corresponding authors needs to increase.

Table 3. Proportion of corresponding authors (CAs) between China and other 5 big nations in 2007 and 2011

<table>
<thead>
<tr>
<th>Co-nations</th>
<th>2007</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Co-paper</td>
<td>CAs</td>
</tr>
<tr>
<td>USA</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>4032</td>
<td>4032</td>
</tr>
<tr>
<td>JAPAN</td>
<td>2684</td>
<td>1363</td>
</tr>
<tr>
<td></td>
<td>3915</td>
<td>2016</td>
</tr>
<tr>
<td>UK</td>
<td>1992</td>
<td>1059</td>
</tr>
<tr>
<td></td>
<td>3716</td>
<td>1985</td>
</tr>
<tr>
<td>GERMANY</td>
<td>1683</td>
<td>793</td>
</tr>
<tr>
<td></td>
<td>2861</td>
<td>1288</td>
</tr>
<tr>
<td>FRANCE</td>
<td>1054</td>
<td>466</td>
</tr>
<tr>
<td></td>
<td>1898</td>
<td>793</td>
</tr>
</tbody>
</table>

How does the role differ according to discipline? Based on the SCI 22 subjects/disciplines we determine the proportion of corresponding authors in international collaborative papers for the years 2007 and 2011. The data shown in Table 4 reveal that in 2007, the following disciplines have a proportion of corresponding authors higher than the average (54.0%): agricultural sciences, chemistry, computer science, engineering, environment / ecology, geosciences, materials science, interdisciplinary, physics, plant and animal science. In the following disciplines the opposite is the case: clinical medicine, economics and business, immunology, molecular biology and genetics, neuroscience and behavior, social sciences (general). But in 2011, there are some changes. The following disciplines have a proportion of corresponding authors that is higher than average (56.2%): agricultural science, computer science, engineering, environment/ ecology,
geosciences, materials science, physics, plant and animal science, and space science. Some disciplines in which this proportion was higher than average in 2007 are now below average and vice versa. The collaborative counts of disciplines have doubled from 2007 to 2011; however, in terms of collaborative roles, in particular in terms of playing the leading role some disciplines, such as clinical medicine, immunology, molecular biology and genetics, and neuroscience & behavior, still show room for improvement.

**Table 4. Proportions of corresponding authors (CAs) in collaborative disciplines in 2007 and 2011**

<table>
<thead>
<tr>
<th>Disciplines</th>
<th>2007 Co-paper</th>
<th>2007 CAs</th>
<th>B/A</th>
<th>2011 Co-paper</th>
<th>2011 CAs</th>
<th>B/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agricultural Sciences</td>
<td>406</td>
<td>228</td>
<td>56.2%</td>
<td>870</td>
<td>556</td>
<td>63.9%</td>
</tr>
<tr>
<td>Biology &amp; Biochemistry</td>
<td>956</td>
<td>492</td>
<td>51.5%</td>
<td>1865</td>
<td>948</td>
<td>50.8%</td>
</tr>
<tr>
<td>Chemistry</td>
<td>2586</td>
<td>1493</td>
<td>57.7%</td>
<td>4678</td>
<td>2614</td>
<td>55.9%</td>
</tr>
<tr>
<td>Clinical Medicine</td>
<td>2246</td>
<td>924</td>
<td>41.1%</td>
<td>5231</td>
<td>2495</td>
<td>47.7%</td>
</tr>
<tr>
<td>Computer Science</td>
<td>634</td>
<td>373</td>
<td>58.8%</td>
<td>1591</td>
<td>1015</td>
<td>63.8%</td>
</tr>
<tr>
<td>Economics &amp; Business</td>
<td>59</td>
<td>24</td>
<td>40.7%</td>
<td>127</td>
<td>68</td>
<td>53.5%</td>
</tr>
<tr>
<td>Engineering</td>
<td>2452</td>
<td>1413</td>
<td>57.6%</td>
<td>5124</td>
<td>3221</td>
<td>62.9%</td>
</tr>
<tr>
<td>Environment/Ecology</td>
<td>786</td>
<td>444</td>
<td>56.5%</td>
<td>1570</td>
<td>995</td>
<td>63.4%</td>
</tr>
<tr>
<td>Geosciences</td>
<td>1109</td>
<td>664</td>
<td>59.9%</td>
<td>1901</td>
<td>1172</td>
<td>61.7%</td>
</tr>
<tr>
<td>Immunology</td>
<td>190</td>
<td>76</td>
<td>40.0%</td>
<td>317</td>
<td>142</td>
<td>44.8%</td>
</tr>
<tr>
<td>Materials Science</td>
<td>1431</td>
<td>872</td>
<td>60.9%</td>
<td>2656</td>
<td>1617</td>
<td>60.9%</td>
</tr>
<tr>
<td>Mathematics</td>
<td>934</td>
<td>476</td>
<td>51.0%</td>
<td>1438</td>
<td>795</td>
<td>55.3%</td>
</tr>
<tr>
<td>Microbiology</td>
<td>347</td>
<td>170</td>
<td>49.0%</td>
<td>651</td>
<td>359</td>
<td>55.1%</td>
</tr>
<tr>
<td>Molecular Biology &amp; Genetics</td>
<td>545</td>
<td>228</td>
<td>41.8%</td>
<td>1141</td>
<td>543</td>
<td>47.6%</td>
</tr>
<tr>
<td>Multidisciplinary</td>
<td>244</td>
<td>133</td>
<td>54.5%</td>
<td>480</td>
<td>238</td>
<td>49.6%</td>
</tr>
<tr>
<td>Neuroscience &amp; Behavior</td>
<td>390</td>
<td>157</td>
<td>40.3%</td>
<td>724</td>
<td>334</td>
<td>46.1%</td>
</tr>
<tr>
<td>Pharmacology &amp; Toxicology</td>
<td>375</td>
<td>195</td>
<td>52.0%</td>
<td>621</td>
<td>319</td>
<td>51.4%</td>
</tr>
<tr>
<td>Physics</td>
<td>3074</td>
<td>1760</td>
<td>57.3%</td>
<td>4607</td>
<td>2596</td>
<td>56.3%</td>
</tr>
<tr>
<td>Plant &amp; Animal Science</td>
<td>1031</td>
<td>588</td>
<td>57.0%</td>
<td>1684</td>
<td>974</td>
<td>57.8%</td>
</tr>
<tr>
<td>Psychiatry/Psychology</td>
<td>80</td>
<td>39</td>
<td>48.8%</td>
<td>194</td>
<td>103</td>
<td>53.1%</td>
</tr>
<tr>
<td>Social Sciences, general</td>
<td>125</td>
<td>55</td>
<td>44.0%</td>
<td>308</td>
<td>155</td>
<td>50.3%</td>
</tr>
<tr>
<td>Space Science</td>
<td>323</td>
<td>171</td>
<td>52.9%</td>
<td>549</td>
<td>309</td>
<td>56.3%</td>
</tr>
</tbody>
</table>

**Overseas Chinese play the role of bridges in international collaboration**

The connecting role of countrymen or former countrymen working in other countries has been studied in several publications (Webster, 2004; Jin et al, 2007; Jonkers & Tijssen, 2008). When it comes to Chinese scientists we will use the term overseas Chinese referring to those scientists who are of Chinese descent, but have joined other nationalities, as well as those Chinese which temporarily work and study in other nations. Among the 38,327 Chinese international collaborative papers in 2011, there
are 25,995 papers including overseas Chinese, about 67.8%.

The proportion of overseas Chinese collaborative papers in 22 disciplines (Table 5) shows that in some disciplines, such as physics (75.1%), interdisciplinary studies (74.2%), immunology (73.5%), materials science (73.5%), molecular biology and genetics (73.2%), and computer science (72.2%), the overseas Chinese are very active, often leading the collaboration. Even in some disciplines with a lower proportion of overseas Chinese collaborative papers, such as plant & animal science (50.4%). In sum, the overseas Chinese have a very strong role as a bridge in international collaboration with mainland China.

### Table 5. Proportion of overseas Chinese collaborative papers in 22 disciplines in 2011

<table>
<thead>
<tr>
<th>Disciplines</th>
<th>2011 total numbers of collaborative papers</th>
<th>2011 numbers of Overseas Chinese papers</th>
<th>proportion of overseas Chinese papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agricultural Sciences</td>
<td>870</td>
<td>492</td>
<td>56.6%</td>
</tr>
<tr>
<td>Biology &amp; Biochemistry</td>
<td>1865</td>
<td>1335</td>
<td>71.6%</td>
</tr>
<tr>
<td>Chemistry</td>
<td>4678</td>
<td>3146</td>
<td>67.3%</td>
</tr>
<tr>
<td>Clinical Medicine</td>
<td>5231</td>
<td>3508</td>
<td>67.1%</td>
</tr>
<tr>
<td>Computer Science</td>
<td>1591</td>
<td>1148</td>
<td>72.2%</td>
</tr>
<tr>
<td>Economics &amp; Business</td>
<td>127</td>
<td>80</td>
<td>63.0%</td>
</tr>
<tr>
<td>Engineering</td>
<td>5124</td>
<td>3589</td>
<td>70.0%</td>
</tr>
<tr>
<td>Environment/Ecology</td>
<td>1570</td>
<td>961</td>
<td>61.2%</td>
</tr>
<tr>
<td>Geosciences</td>
<td>1901</td>
<td>1171</td>
<td>61.6%</td>
</tr>
<tr>
<td>Immunology</td>
<td>317</td>
<td>233</td>
<td>73.5%</td>
</tr>
<tr>
<td>Materials Science</td>
<td>2656</td>
<td>1953</td>
<td>73.5%</td>
</tr>
<tr>
<td>Mathematics</td>
<td>1438</td>
<td>867</td>
<td>60.3%</td>
</tr>
<tr>
<td>Microbiology</td>
<td>651</td>
<td>430</td>
<td>66.1%</td>
</tr>
<tr>
<td>Molecular Biology &amp; Genetics</td>
<td>1141</td>
<td>835</td>
<td>73.2%</td>
</tr>
<tr>
<td>Multidisciplinary</td>
<td>480</td>
<td>356</td>
<td>74.2%</td>
</tr>
<tr>
<td>Neuroscience &amp; Behavior</td>
<td>724</td>
<td>515</td>
<td>71.1%</td>
</tr>
<tr>
<td>Pharmacology &amp; Toxicology</td>
<td>621</td>
<td>442</td>
<td>71.2%</td>
</tr>
<tr>
<td>Physics</td>
<td>4607</td>
<td>3459</td>
<td>75.1%</td>
</tr>
<tr>
<td>Plant &amp; Animal Science</td>
<td>1684</td>
<td>851</td>
<td>50.5%</td>
</tr>
<tr>
<td>Psychiatry/Psychology</td>
<td>194</td>
<td>102</td>
<td>52.6%</td>
</tr>
<tr>
<td>Social Sciences, general</td>
<td>308</td>
<td>172</td>
<td>55.8%</td>
</tr>
<tr>
<td>Space Science</td>
<td>549</td>
<td>350</td>
<td>63.8%</td>
</tr>
</tbody>
</table>

### Conclusion and discussion

This article analyzed the trends of international collaboration in science and technology using six leading nations (the United States, China, Japan, Germany, UK and France) as case studies. These data retrieved from the WoS show that the numbers of collaborative papers in five nations have grown much faster than the numbers of
non-collaborative ones. This means that the major scientific and technological nations focus on their own research capability while further promoting science and technology through international collaboration. The numbers of collaborative nations are continuously growing all over the world, but especially in China.

As a not-yet-fully-developed nation, China is playing an increasingly important role in international collaboration. Numbers of collaborative or non-collaborative papers have experienced a breakthrough growth. China’s international collaboration in science and technology has obvious disciplinary preferences, focusing mainly on traditional fundamental disciplines. Overseas Chinese play an extremely important role in mainland China’s international collaboration in science and technology, providing an optimal channel to overcome linguistic and cultural barriers. It is no surprise that China's S&T policies support and facilitate such collaborations.

Collaboration is an optimal strategy to promote science and technology

Why does the data show sharply increasing patterns of international collaboration in recent years?

In researchers’ level, collaboration is intrinsic needs in their careers. Everyone want to publish instead of perish, especially to publish in high level journal, such as *SCIENCE* or *NATURE*. Many data have verified the fact that partnering with collaborators anywhere can help them do the research, and even more importantly, get their papers accepted, especially partnering the prestigious professors. In earlier stage, the collaboration has taken place in colleagues, supervisor and students, even different institutes. With the support of IT, the collaboration gradually spread more remote distance, such as different states or nations. It can be called bottom-up pattern, which is driven by the researcher himself and is a self-organized process. The limitations of this pattern are stochastic and non-planned, but it is a critical factor to promote the sharp increment of collaboration, whether the local collaboration or international one.

International collaboration in science and technology besides being an effective way to boost productivity also promotes prosperity in the world. No nation, developing or developed, has enough resources to match its needs in science and technology. Therefore international collaboration has become an important aspect in scientific policies. We even claim that it is the optimal path to acquire needed resources. This is obvious in the case of Big Science projects such as CERN’s Large Hadron Collider (LHC).

Developed nations often have first-mover advantages in research and have enough resources to attract talented, excellent intellectuals from all over the world to participate in international collaboration. Due to their strength, they can be more selective with respect to the nature of their collaboration. In order to reach the research front of a discipline, developed nations have paid a lot of attention to increase their own scientific and technological capabilities, while promoting international collaboration and formulating appropriate international collaboration strategies.
Due to limitations in scientific and technological conditions it is impossible for developing or non-developed nations alike to carry out comprehensive research in all disciplines. Hence they have to formulate suitable strategies focusing on their best disciplines on the one hand, and promoting their other disciplines through international collaboration. Generally, developing nations often have an advantage in the number of human personnel, so that they have manpower to offer when collaborating with developed nations. Such collaborations may accelerate the development of science and technology in the world.

Therefore, international collaboration in science and technology is an optimal way to reallocate scientific and technological resources on a global scale. Every country sees international collaboration as a tool to develop its own R&D strength. This collaboration can be call top-down pattern, which is an optimal strategy driven by government policies. Not only the developing countries, such as China, but some developed ones, such as UK, have formulated some planning to promote international collaboration in science and technology. The impressive characteristics of this pattern are the target-oriented and the planned, and it can be used to interpret the preferred collaborative disciplines, institutions, and countries.

20 years ago or earlier, it is difficult to seek a collaborator in remote distance, especially different countries. It is the main thing that the advent of the internet has given more convenient condition to promote collaboration during the past two decades, especially the international cooperation. With the development of the Internet, scientific and technological collaboration in an international context has become easier. Gone are the days of hand written letters, difficult intercontinental telephone talks and an occasional long travel. Nowadays colleagues living on the other side of the globe are as close for discussions as a colleague working two stages higher in the building.

All of these, not only the needs of single researchers but the governmental policies and context changes of information techniques, together promote the international collaboration in science and technology. The bottom-up practices of collaborative needs in scientists have accelerated the increment of collaborative papers and settled the micro-foundation of collaboration, the top-down collaborative policies formulated by the government have facilitated the policies context, and the information techniques, particularly the internet, have eliminated the spatial barriers of communication and collaboration in long distances. The prosperities of international collaboration are due to the harmonious function of above factors.

*Ethnic ties are a bridge to optimize collaborative channels for China*

For developing nations, it is an important issue to seek an international collaborative partner in science and technology. Yet there exist many barriers to collaboration, such as culture, language, funds and policies. It is known that building on ethnic ties is an optimal channel to eliminate such barriers (Jin et al., 2007).

The ‘Overseas Chinese phenomenon’ (Jin et al., 2007) in China’s international
collaboration has developed from a deep social basis over many years. Chinese scientists leaving China, permanently or temporarily, are embedded in traditional Chinese culture, writing the same language, and having close social relations with China. These factors form a strong social basis binding overseas Chinese to a country with which they share many characteristics. This social basis is also where the opportunity lays for overseas Chinese to carry on international collaboration with China. From the foreign country’s perspective, the presence of a significant community of ethnic Chinese makes it easier to collaborate with scientists in mainland China. Indeed, Jonkers and Tijssen (2008) conclude that while countries may lose human capital when Chinese or other ethnic groups return to their homeland (a phenomenon referred to as “return brain drain”) they may also gain in terms of scientific linkages.

Furthermore, the government of mainland China has formulated several specific policies regarding overseas Chinese, such as *The Recruitment Program of Global Experts*, *Introducing Talent intellectuals Planning*, *Yangtze River Scholars Program*, *Thousand Talents program*, and so on. These programs attract outstanding overseas Chinese persuading them to return to their homeland.

Such ethnic ties are of a homophilic nature (McPherson et al., 2001), referring to the phenomenon that similar groups have more frequent contact than un-similar ones. On the one hand, we notice that overseas Chinese scientists establish emotional contact through family, relatives, teacher-student relationship and former collaboration, making future collaboration easier; and on the other hand, based on ethnic ties, this collaboration can enhance the prestige of participants in mainland China. In fact, this phenomenon may not only exist for Chinese, but also for other ethnic groups such as Indians and Koreans.

**References**


BIBLIOMETRIC ANALYSIS AND RANKING OF LIBRARY AND INFORMATION SCIENCE (LIS) RESEARCH AND PUBLICATIONS IN AFRICA

Okon E. Ani1 Eucharia Okwueze2

1 anioedet@yahoo.com
University Library, University of Calabar, Calabar, Nigeria

2 okwuezeucharia@gmail.com
Department of Library and Information Science, University of Calabar, Calabar, Nigeria.

Introduction
Bibliometric analysis is a method that is used to describe patterns of publication and national and international strengths and biases in areas of research within a given field or body of literature such as library and information science (LIS) (Aharony, 2011). Johnson (2011) explained that many bibliometric studies enumerate the extent and scope of scientific communication amongst scholars/researchers. According to Ocholla and Ocholla (2007), bibliometric studies are widely used to inform policies and decisions in political, economic, social and technological domains affecting information flow and use pattern within, between and outside institutions and countries. Thus, bibliometric studies are essentially used in evaluation of research and publications in all areas of human endeavors to support decision making by policy makers in governments, organizations/institutions (universities), academic departments and individual scholars/researchers. Furthermore, bibliometric analysis is significantly used in the ranking of research and publications (Meho and Spurgin, 2005) and this has become popular especially in the universities and other academic and research institutes with emergence of different rankings organizations within the past one or two decades (Ani, 2015). Barik and Jena (2016) opined that bibliometric study is focus on evaluation of research performance of authors, institutions/universities, countries or regions in variety of disciplines. Publications counts and citation analysis are two major methods that are used in bibliometric studies. Ocholla and Ocholla (2007) described citation analysis as a tool that is used as quantitative measure of the quality of research and publications.

LIS is a discipline that is fast growing in terms of research and publication globally. Although bibliometric analysis and ranking studies in LIS have a long history in developed countries (Meho & Spurgin, 2005); in Africa, bibliometric studies are limited and insignificant in Africa in spite of the widely applications of these studies by policy makers to support research and publications in line with the emerging global rankings of universities (Ocholla & Ocholla, 2007). Onyancha (2007) vividly captured the poor state of LIS research in Africa by expressing the fear that there is general concern among LIS scholars concerning the growth and development of LIS research in Africa in relation to global perspective. Thus, this present study is intended to fill the knowledge gap in bibliometric studies and would therefore expand the scope of literature in LIS in Africa.

Objectives of the Study
The study was guided by the following objectives:
1. To rank LIS research and publications in African countries in global perspective;
2. To rank LIS research and publications in Africa by country, and
3. To determine trends in LIS research and publications in Africa.

Research Method
The Web of Science databases published by Thomson Reuters were used for the study. The study was limited to the used of Social Science Citation Index (SSCI) and Arts and Humanities Citation Index (A&HCI). The data for the study were obtained by conducting an advanced search using Web of Science Category, WC=Information Science and Library Science, with articles as the document types, all languages in terms of selection of language of publication of the journal articles. The period of the study was limited to 2010-2016 to depict current state of research globally and in Africa in particular. The data obtained were analyzed globally and by African countries based on the objectives of the study.

Results and Discussion
The results of the study are presented and discussed in this section based on the objectives of the study. Ranking of LIS research and publications in Africa in global perspective
The results of the study showing ranking of LIS research and publications in Africa in global perspective are presented in table 1. The results revealed that Africa is not represented among the top 10 countries globally in LIS research and publication in all bibliometric indicators that are used in this study (publication counts and citation...
analysis: total citations, citations per article and h-index). Earlier study by Sin (2005) has confirmed this finding that Africa is not represented among the top 10 productive countries in LIS research and publication. Analysis of the results showed that USA is significantly leading globally in LIS research and Africa is only represented in the global ranking among the 20 top countries by South Africa in LIS research and publication in number 19 positions with publication counts of 364.

Table 1: Top 20 globally ranked countries in LIS publications

<table>
<thead>
<tr>
<th>S N</th>
<th>Countries</th>
<th>Total Articles</th>
<th>Total Citations</th>
<th>Citations per article</th>
<th>H-Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>USA</td>
<td>9.782</td>
<td>68,176</td>
<td>6.97</td>
<td>73</td>
</tr>
<tr>
<td>2</td>
<td>China</td>
<td>2.025</td>
<td>12,664</td>
<td>6.25</td>
<td>38</td>
</tr>
<tr>
<td>3</td>
<td>England</td>
<td>1.969</td>
<td>13,471</td>
<td>8.84</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>Spain</td>
<td>1.848</td>
<td>8,171</td>
<td>4.42</td>
<td>33</td>
</tr>
<tr>
<td>5</td>
<td>Canada</td>
<td>1.421</td>
<td>9,547</td>
<td>6.72</td>
<td>38</td>
</tr>
<tr>
<td>6</td>
<td>Australia</td>
<td>1.246</td>
<td>9,649</td>
<td>5.57</td>
<td>31</td>
</tr>
<tr>
<td>7</td>
<td>Germany</td>
<td>1.000</td>
<td>6,631</td>
<td>6.63</td>
<td>53</td>
</tr>
<tr>
<td>8</td>
<td>Taiwan</td>
<td>869</td>
<td>5,702</td>
<td>6.53</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>South Korea</td>
<td>866</td>
<td>5,455</td>
<td>6.30</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>Netherlands</td>
<td>832</td>
<td>8.721</td>
<td>10.48</td>
<td>39</td>
</tr>
<tr>
<td>11</td>
<td>Brazil</td>
<td>740</td>
<td>1,077</td>
<td>1.64</td>
<td>13</td>
</tr>
<tr>
<td>12</td>
<td>Italy</td>
<td>524</td>
<td>3,108</td>
<td>5.93</td>
<td>23</td>
</tr>
<tr>
<td>13</td>
<td>France</td>
<td>508</td>
<td>2,432</td>
<td>4.79</td>
<td>21</td>
</tr>
<tr>
<td>14</td>
<td>Belgium</td>
<td>410</td>
<td>2,848</td>
<td>6.95</td>
<td>24</td>
</tr>
<tr>
<td>15</td>
<td>Sweden</td>
<td>395</td>
<td>2,386</td>
<td>6.04</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>India</td>
<td>386</td>
<td>1,385</td>
<td>3.59</td>
<td>15</td>
</tr>
<tr>
<td>17</td>
<td>Singapore</td>
<td>381</td>
<td>2,570</td>
<td>6.75</td>
<td>24</td>
</tr>
<tr>
<td>18</td>
<td>Finland</td>
<td>374</td>
<td>2,730</td>
<td>7.30</td>
<td>24</td>
</tr>
<tr>
<td>19</td>
<td>South Africa</td>
<td>364</td>
<td>1,006</td>
<td>2.76</td>
<td>15</td>
</tr>
<tr>
<td>20</td>
<td>Denmark</td>
<td>295</td>
<td>2,041</td>
<td>6.92</td>
<td>22</td>
</tr>
</tbody>
</table>

Table 2: Ranking of African countries by publication outputs

<table>
<thead>
<tr>
<th>S N</th>
<th>Countries</th>
<th>Total Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>South Africa</td>
<td>364</td>
</tr>
<tr>
<td>2</td>
<td>Niger ia</td>
<td>165</td>
</tr>
<tr>
<td>3</td>
<td>Ghana</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>Kenya</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>Egypt</td>
<td>26</td>
</tr>
<tr>
<td>6</td>
<td>Tanzania</td>
<td>24</td>
</tr>
<tr>
<td>7</td>
<td>Botswana</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>Uganda</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>Tunisia</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>Malawi</td>
<td>11</td>
</tr>
<tr>
<td>11</td>
<td>Benin</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>Morocco</td>
<td>10</td>
</tr>
<tr>
<td>13</td>
<td>Zimbabwe</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>Algeria</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>Ethiopia</td>
<td>8</td>
</tr>
<tr>
<td>16</td>
<td>Senegal</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td>Namibia</td>
<td>6</td>
</tr>
<tr>
<td>18</td>
<td>Zambia</td>
<td>4</td>
</tr>
<tr>
<td>19</td>
<td>Mozambique</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>Total</td>
<td>781</td>
</tr>
</tbody>
</table>

Ranking of LIS research and publication in Africa by country

The results of the study showing the rankings of LIS research by publication counts in Africa are shown in table 2. The results in table 2 revealed that South Africa and Nigeria are the foremost countries in LIS research in term of publication counts. The findings of the study are consistent with previous studies that affirmed South Africa and Nigeria as the leading countries in LIS research in Africa (Aharony, 2011; Onyancha, 2007).

Table 2: Ranking of African countries by publication outputs

<table>
<thead>
<tr>
<th>S N</th>
<th>Countries</th>
<th>Total Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>South Africa</td>
<td>364</td>
</tr>
<tr>
<td>2</td>
<td>Niger ia</td>
<td>165</td>
</tr>
<tr>
<td>3</td>
<td>Ghana</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>Kenya</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>Egypt</td>
<td>26</td>
</tr>
<tr>
<td>6</td>
<td>Tanzania</td>
<td>24</td>
</tr>
<tr>
<td>7</td>
<td>Botswana</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>Uganda</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>Tunisia</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>Malawi</td>
<td>11</td>
</tr>
<tr>
<td>11</td>
<td>Benin</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>Morocco</td>
<td>10</td>
</tr>
<tr>
<td>13</td>
<td>Zimbabwe</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>Algeria</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>Ethiopia</td>
<td>8</td>
</tr>
<tr>
<td>16</td>
<td>Senegal</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td>Namibia</td>
<td>6</td>
</tr>
<tr>
<td>18</td>
<td>Zambia</td>
<td>4</td>
</tr>
<tr>
<td>19</td>
<td>Mozambique</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>Total</td>
<td>781</td>
</tr>
</tbody>
</table>

Ranking of African countries by citation analysis

In table 3, African countries are ranked by citation analysis and the results revealed that South Africa is significantly leading other countries in the number of total citations (1006), followed by Nigeria (229), Kenya (106), Ghana (90) and Uganda (90). South Africa (15) is again leading in h-index, followed by Nigeria (7), while Tunisia (7.58) is the most ranked African country in terms of citations per article and is followed by Ethiopia (7.00).
Table 3: Ranking of African countries in LIS research by citation analyses

<table>
<thead>
<tr>
<th>SN</th>
<th>Countries</th>
<th>Total Citations</th>
<th>Citations per article</th>
<th>H-Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>South Africa</td>
<td>1,006</td>
<td>2.76</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>Nigeria</td>
<td>229</td>
<td>1.39</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>Ghana</td>
<td>90</td>
<td>2.37</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>Kenya</td>
<td>106</td>
<td>3.53</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>Egypt</td>
<td>89</td>
<td>3.42</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>Botswana</td>
<td>74</td>
<td>3.08</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>Tanzania</td>
<td>65</td>
<td>2.71</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>Uganda</td>
<td>90</td>
<td>4.29</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>Tunisia</td>
<td>91</td>
<td>7.58</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>Malawi</td>
<td>62</td>
<td>5.64</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>Benin</td>
<td>19</td>
<td>1.73</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>Morocco</td>
<td>32</td>
<td>3.20</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>Algeria</td>
<td>4</td>
<td>0.44</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>Zimbabwe</td>
<td>7</td>
<td>0.78</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>Ethiopia</td>
<td>56</td>
<td>7.00</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>Senegal</td>
<td>14</td>
<td>2.33</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>Namibia</td>
<td>6</td>
<td>1.00</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>Zambia</td>
<td>12</td>
<td>3.00</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>Mozambique</td>
<td>2</td>
<td>0.67</td>
<td>1</td>
</tr>
</tbody>
</table>

Trends in LIS research and publications in Africa, 2010-2016

The growth, progress and development of different fields of knowledge are usually determined by annual publication outputs in a given field (discipline) over a period of time. It was therefore pertinent to examine the trends in LIS research and publication in the study. The results of the study are presented in figure 1. The results showed an increasing trend in general or overall annual publication outputs in LIS research in Africa; but analysis by individual country using the three most ranked countries (South Africa, Nigeria and Ghana), showed declining and fluctuating trends. In South Africa, there was an increased in annual publication output until a decline was set in 2015-2016. The declining and fluctuating trend in annual publication outputs was apparent in Nigeria. Similar declining and fluctuating trend was also observed in annual publication output in Ghana. In the case of Nigeria, a recent study by Ani, Ngulube and Onyancha (2017) affirmed fluctuating and unpredictable trend in the annual publication output in LIS research in Nigeria.

Conclusion

LIS is a vital discipline that has been developed globally to support national and international development. In view of the importance of LIS in socio-economic development, the field has been transformed from its traditional role of information storage, dissemination among others. However, and in spite of the rapid growth, progress and development in LIS research globally, scholars have decried the poor state of LIS research and publication in Africa (Mchombu, 2002; Ocholla & Ocholla, 2007; Onyancha, 2007). The findings of this study have therefore made significant contributions to the literature of growing concerns on the need to promote LIS research and publication in Africa in line with the global trend. The findings of the study have revealed that Africa is not represented among the top 10 productive countries in the world in LIS research and publication, but is however, represented by South Africa in the 19th position among the top 20 productive countries in the world. Thus, South Africa is one of the leading countries in LIS research and publication in the world and Africa in particular. The rankings of African countries in LIS research and publication, however, revealed that South Africa (364) and Nigeria (165) are significantly leading other African countries in publication outputs. Although, the results of the study indicated general or overall increase in annual publication outputs in LIS research in Africa; the trends in annual publication outputs by the three most productive countries (South Africa, Nigeria, and Ghana) were found to be declining and fluctuating.

References

Aharony, N. 2011. Library and Information Science Research Areas: A Content Analysis of Articles from the Top 10 Journals 2007-


Ternary Co-occurrence Latent Semantic Vector Space Model

Niu Fenggao, Wang Shichang and Zhang Yayu

nfgao@sxu.edu.cn
Shanxi University, college of mathematical science, No. 92, Hollywood Road, Taiyuan, Shanxi (China)

2889859712@qq.com
Shanxi University, college of mathematical science, No. 92, Hollywood Road, Taiyuan, Shanxi (China)

1449188717@qq.com
Shanxi University, college of mathematical science, No. 92, Hollywood Road, Taiyuan, Shanxi (China)

Keyword
Co-occurrence analysis, Ternary co-occurrence, Co-word matrix, Weighted CLSVSM, Clustering

Introduction
Resource aggregation is an effective way to solve the question of too many resources. Literature aggregation can be realized by the clustering of the feature vector of literature. It is very important to use the semantic information to express the vector fully. The most basic vector representation is the vector space model (VSM) (Salton, 1997), this model assumes in-dependency between the vocabulary terms and ignores all the conceptual relations between terms that potentially exist; The Generalized vector space model (GVSM) (Wong, Ziarko&Wong, 1985) excavates the co-occurrence information of words, but it cannot extract the semantic information fully. The semantic vector space model (SVSM) (Song, Liang&Soon, 2014), excessive relies on ontology or the language database. The co-occurrence latent semantic vector space model based on the binary information (CLSVSM) (Niu&Qiu, 2014) is used to proceeding literature aggregation. The result shows that the model is better than the VSM. But we find that researching the binary information only is not comprehensive. As a consequence, we introduce the ternary information.

Ternary CLSVSM
The co-occurrence latent semantic vector space model is mainly realized by mining the latent relationship between terms. Latent relationships include synonyms, antonyms and near-synonyms. This passage is mainly based on the research of binary co-occurrence. Next, we will study the model from the following three aspects: the representation of co-occurrence matrix, the calculation of the relative strength and the weighted CLSVSM.

The Representation of Co-occurrence Matrix
In this passage, we present a ternary co-occurrence layer matrix with algebraic method. We can use this matrix to calculate the co-occurrence frequency. Definition1: Given document sets D and its document - term matrix A (only 0,1), let \( t_j \) be the j-th column of A, we define co-occurrence frequency of the j-th term \( t_j \) and all term pair \( (t_i, t_k) \) is the j-th layer of ternary co-occurrence as follow:

\[
C_j^{(3)} = C_j^{(1)}(t_j) = (\text{diag}(t_j) . A)^T \cdot (\text{diag}(t_j) . A)
\]

Explication: We calculate \( \text{diag}(t_j) . A \) first. In calculation, we use \( \text{diag}(t_j) \) to modify the matrix A.

If the diagonal elements of \( \text{diag}(t_j) \) contain 0, that is, the term \( t_j \) at least in one document does not appear (might as well be k), at this time the k-th row and the k-th column elements are 0. According to the matrix multiplication, the k-th row the elements of the \( \text{diag}(t_j) . A \) are all 0. On the contrary, If the diagonal elements of \( \text{diag}(t_j) \) are all 1, the result of \( \text{diag}(t_j) . A \) is A.

For example:

\[
A = \langle t_1, t_2, t_3, t_4, t_5 \rangle = \begin{bmatrix}
1 & 1 & 0 & 1 & 0 \\
0 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 & 1 \\
1 & 1 & 1 & 0 & 0
\end{bmatrix}
\]

\[
\text{diag}(t_1) . A = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
1 & 1 & 1 & 0 & 1 \\
1 & 1 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]

1825
The Relative Strength of Ternary Co-occurrence

We can also use the following method to calculate the ternary co-occurrence frequency. Let \( c_i, c_{ij}, c_{ijk} \) respectively be the frequency of terms \( t_i, t_j \) and \( t_k \). Let \( C_{ij} \) be the number of documents that including at least one of the three terms. Contrasting the promotion of the addition formula in the probability theory, for the three events A,B,C, let P be the probability of occurrence of events, then we define \( P(A \cup B \cup C) \) as follow:

\[
P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)
\]

We can infer the calculation formula of the ternary co-occurrence frequency:

\[
c_{ijk} = C_{ij} + C_{jk} + C_{ik} + C_{ijk}
\]

Definition 2: Given binary co-occurrence matrix C, we can define the relative strength of ternary co-occurrence \( b_{ijk} \) of any three terms \( t_i, t_j, t_k \) as follow:

\[
b_{ijk} = \frac{c_{ijk}}{\sqrt[3]{c_{ij}c_{jk}c_{ik}}}, i, j, k = 1,2,\ldots, m
\]

The Weighted CLSVSM

In the new CLSVSM, we consider combining with binary and ternary co-occurrence information, and give the ternary greater weight.

Definition 3: Given the binary co-occurrence relative strength matrix B, the ternary co-occurrence matrix \( C^{(3)} \), the j-th floor ternary co-occurrence matrix \( C^{(3)}_j \), the j-th floor ternary co-occurrence relative strength matrix \( B^{(3)}_j \),

\[
C^{(3)}_{ij} = \sum_{t_i} \sum_{t_j} \sum_{t_k} c_{ijk} = (\text{diag}(t_i), A) \cdot (\text{diag}(t_j), A)
\]

We find that the second diagonal elements of \( \text{diag}(t_j) \) is 0, therefore the second row of \( \text{diag}(t_j)A \) is 0. \( C^{(3)} \) is the co-occurrence frequency matrix of the first term \( t_i \) with other terms. Obviously, the matrix is a symmetric matrix. The element of the second row and the third column is the co-occurrence frequency of the terms \( t_1, t_2 \) and \( t_3 \).

We define the new co-occurrence latent semantic vector space model based on binary and ternary as follow:

\[
\phi(d_i) = \tilde{a}_i = (q_{i1}, q_{i2}, \ldots, q_{im}) \in \mathbb{R}^m
\]

\[
q_{ij} = \begin{cases} 
\max_{t_i} \left\{ b_{ij} \right\}, & \text{if } a_{ij} = 1, \\
\max_{t_i} \left\{ b_{ij} \right\} + \frac{1}{3} \left( \max_{t_i} \left\{ b_{ij} \right\} \right), & \text{if } a_{ij} = 0, \text{max}_{t_i} \left\{ b_{ij} \right\} = 0, \\
0, & \text{otherwise}
\end{cases}
\]

Where, \( I_{ai} = \{ j | a_{ij} = 1 \} \) is the indicator sets of \( j \).

The new document - term matrix is defined as:

\[
Q = \begin{pmatrix} 
q_{i1} & q_{i2} & \cdots & q_{im-1} & q_{im} \\
q_{i3} & q_{i4} & \cdots & q_{im-2} & q_{im-1} & q_{im} \\
& & \ddots & \vdots & \ddots & \ddots \\
q_{i,n-2} & q_{i,n-2} & \cdots & q_{im-n+1} & q_{im-n+1} & q_{im-n+1} &q_{im-n+1} & q_{im-n} & q_{im-n} & \cdots & q_{im} \\
q_{i,n} & q_{i,n} & \cdots & q_{im-n+2} & q_{im-n+2} & q_{im-n+2} & q_{im-n+2} & q_{im-n+1} & q_{im-n+1} & q_{im-n+1} & \cdots & q_{im-n+1} & q_{im-n+1} & q_{im-n+1} & q_{im-n+1}
\end{pmatrix}
\]

Conclusion

This paper we researched the ternary CLSVSM and gave the experimental design. We showed that, in the Chinese literature, the clustering effect of the weighted model was better but in the English literature, the two clustering effects were equivalent. In this paper, the representation of the ternary co-occurrence layer matrix can be extended to the multiple co-occurrences, but it should be noted that the frequency of the multiple co-occurrence will decrease rapidly with the increase of multiple numbers. The multiple co-occurrence phrases are also rapidly decreasing. In the subsequent study, we will try to find a way to directly represent the ternary information and explore the scope of the model.

References


Scientometric Study on Technology Promoting Science Development

Xie Cai-xia¹, Zhang Yan¹, Wang Li²

¹. Institute for Science Technology and Society, School of Politics and Public Administration, Henan Normal University, Xinxiang 453002;

². International Programs Office, Henan Normal University, Xinxiang 453002)

Abstract
Based on the review of the research achievements of the relationship between science and technology, using scientometrics and network analysis method, taking the database of CNKI and SCI as data source, this paper studies quantitatively the Promoting effect of scanning tunneling microscopy (STM), the specific new technology, on science development. We research the mechanism of STM permeating each area of science research fields, including the influence of STM on science knowledge structure, STM improving the science research field and the evolutionary dynamic process of hot spots in scientific research under the promoting of STM. We also study the role of STM in promoting other relevant technologies’ growth and development. The aim is to reveal how science and technology promote each other and develop in parallel. This study predicts the emergence and development of the future science research field according to burst detection under the influence of STM.

Key words: Relationship between science and technology; scanning tunneling microscopy (STM); promoting each other; developing in parallel

Conference Topic
Science of science

Review of the relationship of science and technology
The relationship between science and technology has long been one of the most important issues in many disciplines. Especially in this Big Science age, the relationship is becoming more complex and closer, and at the same time their boundary is becoming even more obscure. The correlation between science and technology, science and science, technology and technology should be further studied from a new systematic viewpoint. The relationship between science and technology has become a hot topic for present scholars. For example, many scholars are exploring this relationship on macro-level using qualitative method, such as discussing whether science and technology evolved independently or together (Rip, 1992; Nelson, 2004; Barnes, 1982; Lin, 2014). This chapter explores essence of science and technology.
through analyzing the historical process of science and technology development (Gui, 2009; Guo, Guo & Shi, 2007), studying the mutual association patterns of science and technology, and their common knowledge origins (Cao & Wu, 2013; Narin, Hamilton & Olivastro, 1995), and by researching the relationship between science and technology in the context of modern economic and social development (Ju & Liu, 2004; Huang, 2002).

Contemporary quantitative analysis is another method used in this study. Especially literature quantitative analysis is used frequently in this field. In their long development process, science and technology have interacted, and depended on each other. In this process, science has provided the theoretical foundation for technology development and the intellectual preparation for technological innovation. The needs and advances in technology has provided new research subjects, research tools and exploration methods for science as well (Tan & An, 2007). But meanwhile, science and technology, as two different stages in the whole process of human learning have big differences in categories, presentation forms and evaluation standards. The mainly presentational forms of scientific achievement are academic papers, and technology innovations are mainly carried out as invention patents. That is to say, papers and patents are the knowledge carriers of science and technology achievements respectively. Mutual citation between these documents has not only reflected the inherent relevance of their knowledge, but also shown the direction and process of the knowledge flow. By researching the various indicators, contents and inter-reference features of papers and patents, we can reveal the relationship of science and technology in a quantitative perspective as well as the relation of the mutual penetration, interaction and mutual promotion between basic researches and technical innovations. Through exploring the reference links of papers and patents in specific fields, we can analyze the inherent interaction mechanism of science and technology from a micro-perspective.

The reference between patents and papers is bi-directional. In other words, papers can be cited by patents as well as patents do. To some extent, patents referencing to papers has reflected how research achievements are applied to technological invention. Through analyzing the characteristics and laws of the patents referencing to papers, we can reveal the science research’s effects on technology innovation. The specific indexes used to measure the patents referencing papers mainly contain time distributions, regional distribution, periodicals distribution, as well as subject analysis and subject categories. Carpenter MP and Narin F. applied the analysis of patents referencing to papers in researching the relationship between science and technology earlier (1983). Based on the SUPTO data from 1987 to 1996, Tijssen studied the relationship between science and technology with the analysis methods of social network (2001). Combining patents citations analysis methods with econometric models, Van Looy B found that there was a strong correlation between science and technology (2003). Other scholars explored their relationship’s features through analyzing the patents referencing to papers in specific subjects fields. For example, Narin studied the citation relationship between patents and papers of bioscience, and found that the more advanced the technology was, the further closer to basic sciences it was (Narin & Noma, 1985). Zhao Zhiyun etc, who conducted an empirical study on Chinese biotechnology, found that the basic research and technology was closely linked and the relation between them was becoming closer over time (Zhao & Lei, 2013).
Papers referencing to patents has reflected how technology has affected scientific research. Through the law of paper citing patent, the relationship of basic research to technological innovation can be demonstrated from another perspective, and the impact of technology on science can be measured. For instance, through analyzing the references of SCI papers to USPTO patents, Glanzel W and Meyer M have built internal relations between science and technology and pointed out the significant difference of science and technology’s relevance in different subjects (Glanzel&Meyer,2003). However, by contrast, there are few examples of quantitative research on relation of science to technology, through analyzing the references of papers to patents. The main reason is that citation papers on patents tend to be too few to query, and the shortage of data might statistically affect the credibility of the results of analysis. Yang Zuguo and Chen Hong etc. have explored the status and laws of China's patent technology being used for scientific research by collecting the data of scientific papers citing Chinese patents, and by analyzing them. In their conclusion, they pointed that due to the few sources of statistical data, the statistical analysis conclusion had some limitations (Yang, Chen & Zhu, 1999). According to the data of papers and patents citing each other, Wu Feifei and partners built two relationship networks: science citing technology and technology citing science, based on the citation data of papers and patents citing each other. Combining the relationship networks with social network analysis methods, they analyzed the interaction of science and technology, and identified the science and technology fields with high network centrality. Through the cross-influence between disciplines and technology fields, they explored the transference between scientific knowledge and technological application. But their analysis of the data of papers citing patents is based on basic data used by Glanzel W in 2003 (Wu, Huang & Shi, 2013). The promotion of technology over science has long been recognized, it is well known that technology development promotes science research, yet there is weak quantitative research for this. Application of new technology provides new ways of observation and experiments, offering new materials for science research; this causes scientists to conduct further research and achieve new breakthroughs in basic scientific research. This is one aspect of the technology driving science development. At the same time, technological development provides new contents for science research, and the application of new technologies might also bring several new problems which have to be resolved. And then all those stimulate further science research. In addition, the natural paths of technology improvement indicate its limitations, while these limitations demand that research enlarges its areas and defines new directions. This is another aspect of technology driving science development. Therefore, the development and breakthrough of modern science often occurs after a crucial improvement of technology. Without the application of technology, science will be difficult to develop. For example, the invention and improvement of experiment instruments plays an important role in science progress, and brings the potential for major breakthroughs in science development. In 1995, the United States National Science Foundation organized an academic conference, and renowned biologists, technical experts and relevant managers attended the conference. The main purpose of this conference was to discuss the impact of emerging technologies on bio-science development. At the conference, speakers thought that if the rapid development of bio-science was achieved, new technology must be combined with the bio-science research
In this study, we start with a specific technology, scanning tunneling microscopy (STM), to study the effects of a specific technique for science research using different analysis methods of patents citation in papers. Taking STM as an example, we investigate how this technology has improved science development, and its impact on each subject. In 1981, Gerd Bining, Heinrich Rohrer and their laboratory staffs of the Zurich laboratory in International Business Machines Corporation had successfully developed one novel surface analysis instrument for the first time in the world --the scanning tunneling microscope. The scanning tunneling microscope helped people "see" the atoms on the objects’ surface, and it could even identify approximately 1% of one atom’s square. Because of the important application of this microscope, their emergence has extended into plenty of new fields, nanotechnology included. This new area has rapidly penetrated into the research of each subject and promoted their development. So the scanning tunneling microscope was recognized as one of the top ten science and technology achievements of the 1980’s by the international science community. Therefore, Gerd Bining and Heinrich Rohrer, the inventors of the scanning tunneling microscope, and Ernst Ruska, the electron microscope inventor, shared the physics Nobel Prize in 1986.

Data source and methodology

We used the published STM-related academic papers as samples. The theses of domestic journals were chosen from CNKI database, and the theses of foreign journals were from the SCIE database of Thomson Web of Science. Using "scanning Tunnel microscopy", "tunnel scanning microscope", "STM", "scanning tunneling microscope" as retrieve terms, we searched the CNKI database and found that the earliest literature appeared in 1985. So we set the period between 1985 and 2015 for investigating, and got 2,223 related articles altogether. In this context, we would like to call the four keywords above-mentioned together as "Scanning Tunneling Microscopy". Searching with the retrieve formula of TOPIC= (SCANNING TUNNEL* MICROSCOP*) AND ADDRESS= (PEOPLES r CHINA) in SCI database, we got 2,485 articles in total. As it is shown in Graph 1, the retrieved documents in two databases were distributed over time. It can be seen that CNKI papers presented a wavelike rising trend before 2001 and then declined rapidly. The publication of SCI papers started in 1989 and grew swiftly and violently from 2005. One of the possible reasons is that a great number of papers associated with the scanning tunneling microscope had started contributing to international journals since 2005. Accordingly, in order to make the data mutual-complement, we combined the two databases and analyzed them in this study.
Currently, there are abundant software and technological means used for Scientometrics and Bibliometrics research, and each of them has its own unique style, advantages and disadvantages. We can choose different software functions based on different contents. For instance, the calculation function of UCINET is powerful, and it can output the calculated data and results to software like NETDRAW for Visual display, but UCINET requires the raw data to be presented in matrix format. When applying UCINET for central calculation and visualized network construction of data, firstly we convert data into BIBEXCEL format by using the CITESPACE conversion function. Then we construct the matrix format for UCINET analysis by BIBEXCEL. One distinctive feature of CITESPACE is to explore the dynamic evolution of research progress and frontiers in specific fields. CITESPACE can detect the burst words from numerous titles, abstract, descriptors and identifiers through BURST DETECTION algorithm, and analyze the evolutionary process and progress of science based on the swift changing trends, rather than the frequency of words only. Therefore, we have conducted the research with appropriate technological means depended on the contents in this paper.

The osmosis of STM technology to science research

The analysis of application breadth of STM technology

STM technology is widely applied in interdisciplinary studies, which was deeply rooted in quantum mechanics, soli-state physics, chemical physics, electron physics, mechanical engineering and cybernetics. The subjects scope of this technology application is comprehensive. We analyzed the quantity of achievements of STM being applied in various disciplines. Using the standards of subjects classification of SCI and CNKI database, 2,223 references with the theme of STM were retrieved from CNKI. They belong to 40 different subject areas, of which the key fields are shown in Figure 2. Moreover, the 2,485 papers retrieved from the SCI database belong to 43 fields, of which the main subjects are shown in Figure 3. By contrast, the disciplines of the papers collected by SCI are more dispersive and
more concentrated. 60% of them are on the topics of physics and chemistry. It is demonstrated that the impact of STM varies in the development of different subjects. In terms of these disciplines’ categories, physics, biology and chemistry belong to the recognized discipline of natural science; instrumentation, industrial, engineering, materials science, metallurgy and metalworking, industrial technologies and equipment, organic chemicals, inorganic chemicals, electricity industry belong to the engineering science and technology disciplines; radio-electronics belongs to information science and technology disciplines.

Figure 2. The subjects distribution of STM-related theses from CNKI database

Figure 3. The subjects distribution of STM-related theses from SCI database

The detection about the speed of STM permeating into various disciplines

The invention of new technologies will promote the development of disciplines in many ways. Some of them are revolutionary, while some are mildly progressive. In modern science research, new technologies have become one of the most important forces of science development. We have analyzed the permeation speed of STM into interdisciplinary studies. In other words, we have studied the speed of various disciplines responding to STM technology. Our research is conducted with retrieve data of CNKI database, according to the time of when those theses were published in this database. We use both the invention date of STM technology, and the lag time when various disciplines papers has been firstly published, as the indexes. Therefore, we can make the following definition: for a certain subject A, assuming the time of the first published paper of STM as T, the invention time of STM as $T_0$, the lag time of subject responding as $\Delta T$, then $\Delta T = T - T_0 + 1$. Accordingly, we have calculated the measure results of lag time that various subjects responding to STM technology, and listed the top 14 of the most productive disciplines concerning the published papers in Figure 4. As Figure 4 shown,
the left axis expresses the lag time that various disciplines respond to STM technology. The right axis represents the amount of subjects containing STM papers. As figure displayed, in these 14 disciplines, the fastest one responding to STM is technology instrumentation industry, lag time was 5.3. In April and May of 1985, Dai Daoxuan, Shen Hong have published the paper “scanning tunneling microscope”, “the work law and progress of scanning tunneling microscope”, introducing the concepts and principles of STM to this discipline. Generally, the reaction speed to the technology and the literature volume about this technology are positive correlation. The faster a subject reacts to the invention of STM, the more literature it has. Yet, there are also exceptions, such as metal science and metal craft. The STM technology was introduced and applied to this subject in April 1988, by Yan Qing, on Iron and Steel Research Journal. With the reaction speed ranking in the fourth place of various subjects, its literature volume was ranked seventh. On the contrary, the reaction speed in the field of radio electronics was slow. In February 1989, Chen Yufeng introduced the concept of STM, eight years after the invention of the STM technology. With the response lag time ranking seventh in various disciplines, its literature volume grew rapidly, and was ranked fourth in the total amount of literature of all subjects.

Figure 4. The comparison of response speed to STM technology in various disciplines

Measure of STM technologies on the role in promoting science and technology
Brooks had proposed that if we can make a map out of all the known knowledge nodes, then the map will reflect the structure and relationship between scientific knowledge (Brookes, 1981). Knowledge element is not only the smallest knowledge node, but also a set of the keywords and subject headings (Wen, Hou & Gong, 2007). Professor Liu Zeyuan pointed out that, under certain conditions, a key unit of knowledge may play a genetic role in determining the knowledge evolution and mutation in the specific fields (Liu, 2010). In the following study, using the high frequency key words in data retrieve results as the object, we have discussed the underlying knowledge structure of STM-related scientific findings, and the dynamic evolution of the research focusing on the main area of knowledge. This chapter has revealed the promotion of STM technology for scientific development and the impact on new technologies.
**Analysis of knowledge structure in STM-related research fields**

We have classified the top 60 high frequency key words in the two databases, and have discovered that the keywords of microscope series implements are scanning microscopy, optical microscopy, scanning electron microscope, scanning near-field optical microscope, scanning tunneling microscope, atomic force microscopy, transmission electron microscopy, scanning probe microscopes, microscopy, microtechnique. Following the STM technology invention, people have improved this technology, have enabled it to work under vacuum, low temperature, magnetic fields and other specific conditions. Meanwhile, based on the principle of STM, a series of new scanning probe microscopes have been invented, such as, atomic force microscopy, scanning capacitance microscopy, scanning near-field optical microscopes, scanning near-field acoustic microscope, scanning near-field thermal microscopy, scanning electrochemical microscopy. These new inventions of the microscope provide a powerful tool for exploration on various specialties of substance surfaces and interfaces. And furthermore, they have accelerated the development of the microscopic manufacture technology and microscopic observation technology.

After the invention of the STM technology, nanotechnology, the frontier science studies substance of 0.01~100nm-scale, has introduced to the public. Nanotechnology is using STM for nano-level analyzing on material surfaces, including the manipulation of atoms, molecules and etching on the surface of those. Among all the investigated keywords, words associated with Nano are Nano, nano-materials, Nano-electronics, nanocrystallization, Nano-sciences, nano-technology, nano-science and technology, carbon nanotubes, nano-scale, gold nanoparticles, Nano-fabrication and nano-structures and Nano-devices, Nano-biology, nano-particles, nanowires, carbon nano-materials.

The advent of STM technology dramatically promotes the development of surface science, and offers powerful supporting technology for surface atoms array, single crystal surface structure and surface structure research. It also provides an objective theory basis to the questions of the basic issues, such as “Where is atom? How does it move?”. Keywords associated with the material surface are surface structure, surface topography, surface roughness, matter surfaces, metal, silicon and graphite surfaces; and there are keywords of material structure and scale: single-layer, molecule, single molecule and single atom, hair, Atomic layer, Atomic scale, atomic level. In addition, retrieve related data. Within the SCI database, there are more keywords for characterizing specific substances, such as copper (CU100, CU110, CU111), carbon (C-60, CARBON, CARBON NANOTUBES, CARBON-MONOXYDE), Silicon (SI (001) SI (100), SI (111), SILICON) etc. There are three high frequency names of the data retrieved in CNKI database, one is the famous physicist Richard Feynman, the Prize winner who is the first one to put forward the concept of nano-Nobel. One is a chemist, C-60 discoverer, Smalley. And the third one is the nanotechnology expert Bai chunli, the first man who has invented STM in China. No discussion about the STM technology without referring to the academician Bai’s work in China. In 1988, China's first STM with a micro-computer control, data analysis and image-processing system has been developed, leading by Bai chunli. The research team, led by Bai chunli, has successfully developed China's first atomic force microscope (AFM) later. The successful development of STM and AFM has enabled China to
join the advanced ranks in these fields in the world. In addition to the development of those instruments, Bai Chunli is also engaged in the study on STM application. As one of the pioneer researchers, he has applied STM to nano-research field, and has achieved outstanding results. Following the lead of Bai Chunli and other researchers, the STM has flourished in the area of surface analysis in China. And his work also furthers and deepens the application of STM in scientific research (Zhang, 1991).

Furthermore, we have used keywords to analyze the hot spots evolution and co-occurrence networks. Keywords co-occurrence networks are built based on the co-occurrence strength of keywords. By investigating the networks, it is expected to explore the connections between different areas, and study the knowledge structure in the fields of science. In the keywords co-occurrence networks, each keyword is a node on the network. Network connection represents the concordance between the keywords. Then, we have analyzed the centrality of network nodes, and have revealed the status of every node in the whole knowledge structure network and its relationship with other knowledge units. There are three main indicators of measure centrality: centrality of degree (DEGREE), centrality of mediation (BETWEENNESS), centrality of closeness (CLOSENESS).

The Degree centrality index is to measure the of the relations between nodes on the web. The more nodes one can get connected to on the web, the more connection of this node is. That shows the node is in the centre of the web. Degree dedicates the position and influence of the nodes on the web. Higher Degree means the node is closer to the web’s core position, and more likely to control and affect the activities of other nodes. By contrast, lower Degree means the node is further away from the web’s core position and hardly ever interacts with others. Ones like this has little effect on other nodes. What the Betweenness measures is, to what extent does one node locate in between other nodes’ "middle" of the web. That is to say, to what extent the node is others’ "intermediary". Nodes like this played the role of bridge in the web. Betweenness shows the degree of the rule that one node perform in the circulation of the web. The higher betweenness one node is, the more crucial position it is in the circulation of knowledge and information on the web. Closeness is the sum of the shortest distance from one node to others on the web. Lower closeness means one node can connect to others through a shorter distance and less depends on others. The smaller Closeness expressed the stronger ability for one node to get ride of others’ control, in other words, the node is more dominating.

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Degree</th>
<th>Keywords</th>
<th>Betweenness</th>
<th>Keywords</th>
<th>Closeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>scanning tunneling microscopy</td>
<td>9.0</td>
<td>scanning tunneling microscopy</td>
<td>88.1</td>
<td>scanning tunneling microscopy</td>
<td>65.6</td>
</tr>
<tr>
<td>AFM</td>
<td>1.5</td>
<td>nanotechnology</td>
<td>32.5</td>
<td>nanotechnology</td>
<td>66.1</td>
</tr>
<tr>
<td>nanotechnology</td>
<td>2.2</td>
<td>Optical microscope</td>
<td>17.3</td>
<td>Molecular biology</td>
<td>66.6</td>
</tr>
<tr>
<td>probe scanning microscopy</td>
<td>1.3</td>
<td>Surface topography</td>
<td>15.9</td>
<td>probe scanning microscopy</td>
<td>66.6</td>
</tr>
<tr>
<td>Atomic scale</td>
<td>0.6</td>
<td>Single molecule</td>
<td>10.0</td>
<td>life science</td>
<td>66.7</td>
</tr>
<tr>
<td>Atomic level</td>
<td>0.7</td>
<td>Molecular biology</td>
<td>6.0</td>
<td>Optical microscope</td>
<td>66.7</td>
</tr>
</tbody>
</table>
Chart 1 shows that the calculations of Centrality of top 25 keywords whose various centrality indexes are high. The calculations dedicated that the rank of Centrality’s calculations was different from the rank of the keywords’ frequency. Practically, some keywords with low frequency ranking has a high Centrality ranking. For example, the frequency ranking of the word “microscope” came out in the 9th place, while it ranked third in the calculations of Betweenness. Moreover, there are other keywords like these, which had significant differences in words frequency and Betweennes. Such as the Atomic lattice of carbon atoms, molecule, micro-robots, Micron Technology, organic thin films, small size effect, copper etc. And the rank of all Centrality indexes of various keywords is very diversity. Keywords with three Centrality
indexes all ranked in top 25, such as scanning tunneling microscopy, nanotechnology, optical microscopy and scanning probe microscopy. And the keywords, ranked in top 25, only with the calculations of Degree and Betweenness indexes included surface structure and surface morphology, single molecule, Silicon, nano-materials, Nano- and nano-fabrication. The keywords, ranked in top 25, only with the calculations of Betweenness and Closeness indexes, included biological molecules and biological macromolecules, carbon atom, Micron Technology, micro-robots, small size effect and Atomic lattice. And the keywords, ranked in top 25, only with the calculations of Degree and Closeness indexes, included nanotechnology, probe scanning microscopy, life science and material surface, tunnel effect. As shown in Picture 5, it is the web structure of the top 40 high frequency keywords ranked in Degree relevant graphic, through the visualization function of the UCINET.

Figure 5. The web structure of keywords at high frequency

The analysis of research fields about STM related hot spots

Next, we have analyzed the changes of the research on STM-related hot spots through studying the burst keywords. The burst terms of one theme refer to the words or phrases of this theme, whose frequency changed swiftly in one period. Burst terms is the mark of the sudden booming of the hot topic in the research area, as well as the significant measurement of the research direction (Chen, 2004). Compared to the traditional ways of high frequency words analysis, the burst of keywords frequency is more suitable for exploring the transform of research hot spots, and the new trends of research development. When analyzing the burst terms by detection technology and arithmetic, we have used the BURST DETECTION of CITESPACE. This function measures the deeper changes, basing on the burst information of keywords’ frequency, during a period of time.

Figure 6 illustrated the situation of the top 20 keywords with strongest citation burst. There were ten keywords had direct correlation with Nano, out of the top 20 keywords. In terms of burst strength, the strongest one is Nano technology with the strength up to 19.09; the second is atomic level with the burst strength up to 11.16 in the period of 2005-2009. From the time, the keywords having burst before 1995 mainly consisted of atomic scale, optical microscopy,

<table>
<thead>
<tr>
<th>KEYWORDS</th>
<th>Year</th>
<th>Strength</th>
<th>Begin</th>
<th>End</th>
<th>1985-2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>纳米电子</td>
<td>1985</td>
<td>11.1614</td>
<td>1999</td>
<td>1995</td>
<td></td>
</tr>
<tr>
<td>纳米晶体</td>
<td>1985</td>
<td>10.921</td>
<td>2005</td>
<td>2009</td>
<td></td>
</tr>
<tr>
<td>纳米光电子</td>
<td>1985</td>
<td>7.9625</td>
<td>1985</td>
<td>1995</td>
<td></td>
</tr>
<tr>
<td>&quot;纳米管&quot;</td>
<td>1985</td>
<td>7.2213</td>
<td>2003</td>
<td>2011</td>
<td></td>
</tr>
<tr>
<td>&quot;量子计算&quot;</td>
<td>1985</td>
<td>6.8867</td>
<td>2012</td>
<td>2012</td>
<td></td>
</tr>
<tr>
<td>&quot;云&quot;</td>
<td>1985</td>
<td>6.7332</td>
<td>2011</td>
<td>2015</td>
<td></td>
</tr>
<tr>
<td>&quot;纳米硬度加工&quot;</td>
<td>1985</td>
<td>6.1439</td>
<td>2000</td>
<td>2001</td>
<td></td>
</tr>
<tr>
<td>&quot;激光中的&quot;</td>
<td>1985</td>
<td>5.8584</td>
<td>2000</td>
<td>2001</td>
<td></td>
</tr>
<tr>
<td>&quot;纳米材料&quot;</td>
<td>1985</td>
<td>5.2899</td>
<td>2000</td>
<td>2004</td>
<td></td>
</tr>
<tr>
<td>&quot;纳米传感器&quot;</td>
<td>1985</td>
<td>4.9962</td>
<td>2001</td>
<td>2005</td>
<td></td>
</tr>
<tr>
<td>&quot;量子&quot;</td>
<td>1985</td>
<td>4.8096</td>
<td>2004</td>
<td>2006</td>
<td></td>
</tr>
<tr>
<td>&quot;分子束外延&quot;</td>
<td>1985</td>
<td>4.7871</td>
<td>2010</td>
<td>2015</td>
<td></td>
</tr>
<tr>
<td>&quot;纳米加工&quot;</td>
<td>1985</td>
<td>4.7825</td>
<td>1994</td>
<td>1999</td>
<td></td>
</tr>
<tr>
<td>&quot;光学显微镜&quot;</td>
<td>1985</td>
<td>4.7814</td>
<td>1985</td>
<td>1996</td>
<td></td>
</tr>
<tr>
<td>&quot;纳米&quot;</td>
<td>1985</td>
<td>4.7577</td>
<td>2000</td>
<td>2001</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6. Top 20 KEYWORDS with Strongest Citation Bursts**

**Research Conclusions**

There is a close and complex relationship between science and technology. When it comes to conducting quantitative analysis of the relationship of the two, quite a few scholars have revealed how science supports and promotes technology through the citation analysis of patents to theses, and have discussed how technology has accelerated scientific research by citation analysis of theses to patents. In this text, we started with a specific technology to study the rule of how the new technology -STM- has permeated various disciplines of science research. We have also studied the structure of science knowledge under the impact of this technology and how the dynamic evolutionary progress of scientific research hot spots followed this technology. Therefore, in this text, we have conducted a quantitative study using the data of the documents retrieved from the databases of CNKI and SCI and reached several main conclusions below: Literatures associated with STM first appeared in 1985, and kept developing at a stable and swift speed, referring to the disciplines of natural science, engineering science and technology, and information science and technology. STM permeated various disciplines at different speeds, and different subjects also responded to STM discrepancy. Generally, the reaction rates of each
subject to technology had changed in the same direction as that of STM relevant documents. From what we have known of documents about STM, STM was born in the land of quantum mechanics, solid state physics, chemical physics, electronic physics, mechanical engineering, and Cybernetics. Meanwhile, the new technology promoted the emergence of relevant disciplines and also promoted the advance of new relevant technology. STM advanced the appearance and development of the technology associated with microscopy which included atomic force microscopy, SNOM, scanning probe microscopes, scanning electron microscopes, transmission electron microscopy. At the same time, STM also brought rapid progress of research into nanotechnology and surface science. In this field, it can therefore be seen that science and technology not only affect and promote each other, but have also developed along independent paths.

In the co-occurrence web of keywords’ structure, based on the calculations of web Centrality, in indexes of scanning tunneling microscopy, nanotechnology, scanning probe microscopy, and the three Centrality of optical microscopy ranked at the top. The noteworthy keywords are Atomic lattice, micro-robots, small size effect and biological macromolecules. Though they are sorted out of the keywords list of high frequency, the Betweenness and Closeness of these keywords is ranked in front, occupying an important position in the structure of the keywords co-occurrence web. We conducted the research on the hot spots study fields associated with STM technology and found that the burst keyword were MBE, Graphene, quantum computers in recent years. So we can forecast that research associated with these keywords should be the current research hot spots, science breakthroughs and new technological growth in the future.

References:
Cao Dongming, Wu Junjie. (2013). Discuss the basics and patterns of the relevance of science and technology in the modern and present times according to the pedigree cognitive analysis of phenomenon [J]. Study of natural dialectics, 29(11), 26-31.
Ju Naiqi, Liu Guanjun. (2004). Discuss the essence and demand of the transform of science to technology in the perspective of the relations between the two. *Journal of Henan University*(Social Science column), 44(1), 24-28.


Liu Zeyuan. (2010). The thinking of several problems about Knowledge Graph. *Dalian: Dalian University of Technology*.


Analysis on the Developments of Archival Appraisal Theory

Based on Scientometrics

(1990-2016)

Ying He¹ Kun Tian²

¹heying@mail.tjnu.edu.cn
Tianjin Normal University, Tianjin (China)

²ktian@westfield.ma.edu
Westfield State University, Massachusetts (America)

Abstract
Archival Appraisal Theory (AAT) is a fundamental theory of Archival Science, which has been undergoing substantial changes since 1990s. Using the research methods of Scientometrics and the related Knowledge-Mapping tool, this paper analyzes the citations between AAT journal articles and those from other disciplines, identifies the related AAT research hot spots, and summarizes the developments of AAT from 1990 to 2016. We use the study results to extrapolate the trends of AAT developments in the past almost three decades through a macroscopic view. The discoveries could be used to augment and tune Archival Science in order to meet the emerging demands for more applicable and effective theories and methods for appraising modern archival works.

Conference Topic
Citation and co-citation analysis, Mapping and visualization, Social network analysis, Science of science

Introduction
Archival Appraisal Theory (AAT) had been invented as a result of the developments of modern archival works. The theory itself concerns various concepts such as value, utilization, appraisal, simplification, and benefits in appraising archives (B.L.Craig, 2004). Archival characteristics of files and the characteristics of archival management are reflected in AAT (N.Brubach, 1996). Using the research methods of Scientometrics and the related Knowledge-Mapping tool, analysis on the developments of AAT from 1990 to 2016 is made in this paper. Various AAT developments are identified and summarized, and we discover five trends of AAT developments through a macroscopic view.

Archival Appraisal Theory and Its History of Development

What is Archival Appraisal?
Archival appraisal is the practice to identify and assess the values of archives, in order to identify the valuable ones to hand over to an archival institution and to discard the worthless ones meanwhile. Archival appraisal directly determines whether or not an archive should be preserved or destroyed, and the determination itself is the most important and difficult aspect of archival management. (Duranti, 1994; Williams, 2006; Avery, 2010; Cook, 2011).

Developments of Archival Appraisal Theories
A. “Age determination theory”: In 1901, German archival scholar Mays McConnell put forward the notion that “elderly archives should be respected” and he proceeded to set an arbitrary amount of time
for their preservations before destructions. McConnel, for the first time, also raised the idea that the provenance of an archive should be among the most important appraisal criteria. (Fishbein, 1970; Carnicheal, 2004).

B. “Administrative official determination theory”: In the 1920s, British archival scholar Hilary Jenkinson proposed that the archival workers ought not to be involved in the appraisal and destruction of archives. Instead, they should be done by administrative officials (Finch, 1984).

C. “Function appraisal theory”: In the 1930s, Polish archival scholar Carlin Sharansky proposed that the importance of the producers’ positions and their functions in a governmental hierarchical system should be used to decide the values and retention periods of the archives they produced (Boles, 2005; Harris, 2007).

D. “Archival dual values theory”: In 1956 American archival scholar T. R. Schellenberg, known as “the father of archival appraisal theory”, published his iconic work, the Principle and Technology: Modern Archives, and he proposed that public archives should have two different values: the value to government bodies producing the archive, “the first value”, and the value to the users of other government bodies or individual users, “the second value”. Specifically, the first value reflects an archive’s administrative value, the legal value, the financial value and the technological value, while the second value reflects its evidential value, information value and intelligence value. The administrative officials from the government body producing an archive should be responsible for appraising the archive’s first value, whereas its second value should be determined by archival workers (Francis, 2007).

E. “Utilization determination theory”: In 1960s and 1970s, American archival scholars Fez, Bridgetown, Clifford and Finch proposed that historians’ actual and expected utilizations of archives should be the most important criteria for archival appraisal (Ngulube, 2001; Hurley, 2005).

F. “Social analysis and function appraisal theory”: In early 1980s, German archival scholar Booms advocated that an archive should reflect the social values in its formation period. He also argued that these values should be indirectly determined through understanding of its creators’ functions rather than be directly determined through the study of the social public opinions during the period (Behrnd-Klokt, 2009; Pearce-Moses, 2005).

G. “Literature strategy theory”: In mid-1980s, American archival scholar Seymour Wales advocated that the starting point of archival appraisal shall no longer be the examination of an archive but the analysis of the backgrounds in which the archive was produced, especially as a response to the frequent changes of the archive-producing government bodies in modern society (Kurtz, 2004).

H. “Macro-appraisal strategic theory”: In late 1980s, Canadian archival scholar Terry Cook proposed that a variety of factors should be taken into consideration for archival appraisal such as the social structure, archival formation process, creators and their functions (Cox, 2005; Hensen, 2004).

Data Sources and Research Methodology

Data Sources

The Institute for Scientific Information (ISI) started to publish Science Citation Index (SCI) in 1964, and ISI brought an integrated system of academic information resources named “ISI Web of Knowledge”. The core of this system is “Web of Science” (WOS) through which ISI’s three most famous citation databases, Science Citation Index Expanded (the SCIE), Social Science Citation Index (SSCI) and Arts & Humanities Citation Index (A&HCI), can be directly accessed. These three citation databases cover the most important and influential research achievements in the world.

The journal articles from the above three databases of WOS were accessed with a default time span from January 1st 1990 to November 21st of 2016. The information retrieval was centered on the topic of “appraisal”, and the query was “Subject = appraisal or appraise, publication time = 1990-2016, select Social Sciences Citation Index (SSCI) and Arts & Humanities Citation Index (A&HCI), Lemmatization = on, documental type = article”. 20720 articles were retrieved as a result. However, many of the journals articles containing the keywords were irrelevant to archival appraisal research.
For example, an article from an archeology journal containing the keyword “appraisal” may only mean to describe and evaluate the new discoveries from an archeological excavation site and hence does not contribute to archival appraisal research. In order to eliminate such noises, the retrieval results were refined, and eventually 14470 articles were selected as source literatures.

Research Methodology

The interactions between different disciplines are reflected in the citations linking their articles. An article from one discipline citing the research results or study methods introduced by the articles from another discipline shows that the latter discipline informs the first one, which may use the cited results and methods to broaden and/or deepen its own research.

Such analysis is called Citation Analysis, a research method frequently employed in Scientometrics. Making use of graph theory, fuzzy sets, statistics and logical methods such as inference and induction, it analyzes citations on research objects, such as academic journals, conference papers and authors in order to forecast and evaluate the trends of scientific developments.

Knowledge-Mapping is the graphical display of the development processes and the structural relationships of scientific knowledge (Wasserman, 1994; Waaijer etc, 2011; Chen etc, 2002; Anderson, 2008). It makes use of visualization technologies to sort out and describe the relationships between scientific knowledge and the related patterns of knowledge evolution (Newman, 2001; Yan etc, 2012). The developments of AAT in the last thirty years are visualized using VOSviewer1.4, the Knowledge-Mapping drawing tool, in this paper.

VOSviewer, written in the Java programming language, runs on most operating system platforms (Eck, 2010, 2011). It is the primary tool for analyzing a Scientometrics networks. The program can be used to create maps of publications, authors, or journals based on a co-citation network or to create maps of keywords based on a co-occurrence network. The maps are created using the VOS mapping technique and the VOS clustering technique. VOSviewer can be used to view and explore the maps as well. It can show a map in various different ways, with each emphasizing a different aspect of the map.

Results

Analysis 1: Contributions of AAT Related Research Articles by Countries

Table 1 shows that based on the total number of AAT related research articles produced by countries, USA and UK were the ones that are the most active in AAT related research with them together being responsible for more than three fifths of all the articles published during the period. Canada and Australia are also noticeable countries that had highly active AAT related research with each of them accounting for more than 8% of the articles published during the period.

<table>
<thead>
<tr>
<th>Countries/Areas</th>
<th>Number of AAT Related Research Articles</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>6565</td>
<td>33.44%</td>
</tr>
<tr>
<td>UK</td>
<td>6082</td>
<td>30.98%</td>
</tr>
<tr>
<td>CANADA</td>
<td>1907</td>
<td>9.71%</td>
</tr>
<tr>
<td>AUSTRALIA</td>
<td>1635</td>
<td>8.33%</td>
</tr>
<tr>
<td>GERMANY</td>
<td>1146</td>
<td>5.84%</td>
</tr>
<tr>
<td>NETHERLANDS</td>
<td>1099</td>
<td>5.60%</td>
</tr>
<tr>
<td>PEOPLES R CHINA</td>
<td>633</td>
<td>3.23%</td>
</tr>
<tr>
<td>ITALY</td>
<td>611</td>
<td>3.11%</td>
</tr>
<tr>
<td>FRANCE</td>
<td>608</td>
<td>3.10%</td>
</tr>
<tr>
<td>SWITZERLAND</td>
<td>566</td>
<td>2.88%</td>
</tr>
<tr>
<td>SPAIN</td>
<td>488</td>
<td>2.49%</td>
</tr>
<tr>
<td>BELGIUM</td>
<td>415</td>
<td>2.11%</td>
</tr>
<tr>
<td>SWEDEN</td>
<td>355</td>
<td>1.81%</td>
</tr>
</tbody>
</table>
Table 2 lists the frequency statistics of the disciplines citing AAT. It shows that AAT has been cited by many disciplines and thus it has impacts on those disciplines by varying degrees. Figure 1 is a mapping of the relationships among disciplines having cited AAT. The size of a ball on the map is proportional to the number of instances of citing AAT in the related discipline. Figure 2 is the density map of disciplines having cited AAT. Figure 1 and Figure 2 show that the top two disciplines (that cited AAT the most) are Business & Economy and Psychology, which are followed by Social Science, Health Care Sciences & Services, Environmental Sciences & Ecology, etc.

Table 2. Frequencies of Disciplines citing Archival Appraisal Theory (frequency>800)

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Discipline</th>
<th>Frequency</th>
<th>Discipline</th>
</tr>
</thead>
<tbody>
<tr>
<td>17418</td>
<td>Business &amp; Economics</td>
<td></td>
<td>Communication</td>
</tr>
<tr>
<td>14519</td>
<td>Psychology</td>
<td></td>
<td>Information Science &amp; Library Science</td>
</tr>
<tr>
<td>3276</td>
<td>Social Sciences - Other Topics</td>
<td></td>
<td>Family Studies</td>
</tr>
<tr>
<td>2410</td>
<td>Health Care Sciences &amp; Services</td>
<td></td>
<td>Neurosciences &amp; Neurology</td>
</tr>
<tr>
<td>2127</td>
<td>Environmental Sciences &amp; Ecology</td>
<td>1077</td>
<td>Urban Studies</td>
</tr>
<tr>
<td>2025</td>
<td>Public, Environmental &amp; Occupational Health</td>
<td>996</td>
<td>Psychiatry</td>
</tr>
<tr>
<td>1964</td>
<td>Computer Science</td>
<td>995</td>
<td>Social Work</td>
</tr>
<tr>
<td>1809</td>
<td>Public Administration</td>
<td>944</td>
<td>Sociology</td>
</tr>
<tr>
<td>1695</td>
<td>Engineering</td>
<td>873</td>
<td>Education &amp; Educational Research</td>
</tr>
<tr>
<td>1421</td>
<td>Operations Research &amp; Management Science</td>
<td>835</td>
<td>Government &amp; Law</td>
</tr>
<tr>
<td>1223</td>
<td>Management Science</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sport Sciences</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The knowledge mapping of research focuses from the disciplines citing AAT

A keyword reveals the research focus of a scientific research, and the higher frequency one keyword has, the more attention its research focus received. Table 3 lists the frequencies of specifically cited AAT research focuses from the disciplines having cited AAT. Figure 3 is the mapping of relationships among research focuses. Figure 4 is the density map of research focuses from disciplines citing AAT. According to Table 3, Figure 3 and 4, the No.1 research focus is PERFORMANCE, at a frequency of 6092, which is commonly found in the disciplines of Business & Economics, Human Resource Management, and Psychology. It follows that AAT has been applied to management evaluation theories, where PERFORMANCE is a high frequency word. The second research focus is MODEL, which has been found in many disciplines, at a frequency of 4661, and those disciplines normally took advantage of AAT to build mathematic models in order to solve practical problems. The third research focus is BEHAVIOR that has been found in Psychology, at a frequency of 3040, and it follows that AAT have been employed by psychology to study behaviors.
Figure 1. The mapping of relationship between disciplines citing AAT

Figure 2. The density map of disciplines citing AAT
<table>
<thead>
<tr>
<th>Frequency</th>
<th>Keyword</th>
<th>The Main Application Disciplines</th>
</tr>
</thead>
<tbody>
<tr>
<td>6092</td>
<td>PERFORMANCE</td>
<td>Business &amp; Economics, Human Resource Management, Psychology</td>
</tr>
<tr>
<td>4661</td>
<td>MODEL</td>
<td>Many disciplines</td>
</tr>
<tr>
<td>3040</td>
<td>BEHAVIOR</td>
<td>Psychology</td>
</tr>
<tr>
<td>2733</td>
<td>SATISFACTION</td>
<td>Information Science &amp; Library Science, Education &amp; Educational Research</td>
</tr>
<tr>
<td>2533</td>
<td>MANAGEMENT</td>
<td>Management</td>
</tr>
<tr>
<td>2126</td>
<td>IMPACT</td>
<td>Environmental Science</td>
</tr>
<tr>
<td>1893</td>
<td>PERCEPTIONS</td>
<td>Psychology</td>
</tr>
<tr>
<td>1718</td>
<td>PERSPECTIVE</td>
<td>Business &amp; Economics, Management</td>
</tr>
<tr>
<td>1628</td>
<td>INFORMATION</td>
<td>Information Science &amp; Library Science, Computer Science</td>
</tr>
<tr>
<td>1622</td>
<td>WORK</td>
<td>Information Science &amp; Library Science, Archival Science, Psychology, Management</td>
</tr>
<tr>
<td>1586</td>
<td>META ANALYSIS</td>
<td>Medical Science, Psychiatry, Psychology</td>
</tr>
<tr>
<td>1439</td>
<td>ORGANIZATIONS</td>
<td>Sociology, Management</td>
</tr>
<tr>
<td>1360</td>
<td>DECISION-MAKING</td>
<td>Business &amp; Economics, Computer Science</td>
</tr>
<tr>
<td>1326</td>
<td>APPRAISAL</td>
<td>Business &amp; Economics, Archival Science</td>
</tr>
<tr>
<td>1309</td>
<td>HEALTH</td>
<td>Medical Science, Psychology</td>
</tr>
<tr>
<td>1177</td>
<td>STRESS</td>
<td>Medical Science, Psychiatry, Psychology</td>
</tr>
<tr>
<td>1155</td>
<td>VALIDITY</td>
<td>Psychology, Psychiatry, Medical Science, Computer Science, Business Economics, Architecture</td>
</tr>
<tr>
<td>1090</td>
<td>CONSEQUENCES</td>
<td>Law</td>
</tr>
<tr>
<td>1078</td>
<td>PERSONALITY</td>
<td>Psychology, Education &amp; Educational Research, Psychiatry, Law</td>
</tr>
<tr>
<td>1062</td>
<td>QUALITY</td>
<td>Education, Environmental Science, Information Science &amp; Library Science</td>
</tr>
<tr>
<td>1060</td>
<td>OUTCOMES</td>
<td>Medical Science, Hygiene, Management</td>
</tr>
<tr>
<td>996</td>
<td>UNITED-STATES</td>
<td>uncertain</td>
</tr>
<tr>
<td>988</td>
<td>CARE</td>
<td>Medical Science, Medical and Health, Hygiene, Information Science &amp; Library Science</td>
</tr>
<tr>
<td>981</td>
<td>SOCIAL SUPPORT</td>
<td>Psychology, Education &amp; Educational Research, Psychiatry, Medical Science</td>
</tr>
<tr>
<td>977</td>
<td>RATINGS</td>
<td>Finance, Macroeconomic Management and Sustainable Development</td>
</tr>
<tr>
<td>971</td>
<td>WOMEN</td>
<td>Sociology</td>
</tr>
<tr>
<td>962</td>
<td>ATTITUDES</td>
<td>Medical Science, Education, Psychology</td>
</tr>
<tr>
<td>940</td>
<td>STRATEGIES</td>
<td>Business Economics, Education &amp; Educational Research, Psychology</td>
</tr>
<tr>
<td>933</td>
<td>QUALITY-OF-LIFE</td>
<td>Medical Science</td>
</tr>
<tr>
<td>933</td>
<td>SYSTEMS</td>
<td>Computer Science</td>
</tr>
<tr>
<td>925</td>
<td>RISK</td>
<td>Finance, Medical Science, Business Economics</td>
</tr>
<tr>
<td>812</td>
<td>COMMITMENT</td>
<td>Management, Psychology, Business &amp; Economics, Sociology and Statistics, Law</td>
</tr>
<tr>
<td>803</td>
<td>DETERMINANTS</td>
<td>Business Economics, Macroeconomic Management and Sustainable Development</td>
</tr>
<tr>
<td>803</td>
<td>RESPONSES</td>
<td>Medical Science</td>
</tr>
</tbody>
</table>
Figure 3. The mapping of relationship between research focuses from disciplines citing AAT

Figure 4. The density map of research focuses from disciplines citing AAT
Analysis 3: The Citation Analysis of Archival Appraisal Theory (Citing)

The Knowledge-Mapping of disciplines cited by AAT

Table 4 is the frequency statistics of disciplines cited by AAT, and Figure 5 and 6 are respectively the mapping of relationship among disciplines cited by AAT and the density map of these disciplines. These maps show that AAT has referenced and borrowed ideas and research results from other disciplines, such as Business & Economics, Psychology, Social Sciences, Public Administration, Information Science & Library Science, and History. Since 1990s, these disciplines have become important knowledge sources for AAT, impacting its developments.

Table 4. The frequency statistics of disciplines cited by AAT (frequency>100)

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Disciplines</th>
</tr>
</thead>
<tbody>
<tr>
<td>2371</td>
<td>Business &amp; Economics</td>
</tr>
<tr>
<td>968</td>
<td>Psychology</td>
</tr>
<tr>
<td>683</td>
<td>Social Sciences - Other Topics</td>
</tr>
<tr>
<td>559</td>
<td>Public Administration</td>
</tr>
<tr>
<td>292</td>
<td>Information Science &amp; Library Science</td>
</tr>
<tr>
<td>267</td>
<td>History</td>
</tr>
<tr>
<td>246</td>
<td>Family Studies</td>
</tr>
<tr>
<td>219</td>
<td>Social Work</td>
</tr>
<tr>
<td>218</td>
<td>Communication</td>
</tr>
<tr>
<td>214</td>
<td>Health Care Sciences &amp; Services</td>
</tr>
<tr>
<td>193</td>
<td>Environmental Sciences &amp; Ecology</td>
</tr>
<tr>
<td>178</td>
<td>Urban Studies</td>
</tr>
<tr>
<td>171</td>
<td>Computer Science</td>
</tr>
<tr>
<td>151</td>
<td>Government &amp; Law</td>
</tr>
<tr>
<td>148</td>
<td>Philosophy</td>
</tr>
<tr>
<td>147</td>
<td>Public, Environmental &amp; Occupational Health</td>
</tr>
<tr>
<td>120</td>
<td>Anthropology</td>
</tr>
<tr>
<td>120</td>
<td>Sport Sciences</td>
</tr>
<tr>
<td>116</td>
<td>Operations Research &amp; Management Science</td>
</tr>
<tr>
<td>115</td>
<td>Sociology</td>
</tr>
<tr>
<td>100</td>
<td>Medical Informatics</td>
</tr>
</tbody>
</table>

The Knowledge-Mapping of research focuses from disciplines cited by AAT

Table 5 is the frequency statistics of research focuses from disciplines cited by AAT. These research focuses are almost the same as those in Table 3, but there are some differences in the ranking of the focuses. The data in the table suggests that the development of AAT is closely related to the developments of other disciplines with AAT continuously interacting with other disciplines to improve itself to meet their needs. Figure 7 and 8 are respectively the mapping of relationships between research focuses from disciplines cited by AAT and the density map of these research focuses.
Figure 5. The mapping of relationship between disciplines cited by AAT

Figure 6. The density map of disciplines cited by AAT
Table 5. The frequency statistics of research focuses from disciplines cited by AAT

(Frequencies > 60)

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Keyword</th>
<th>The Main Application Disciplines</th>
</tr>
</thead>
<tbody>
<tr>
<td>529</td>
<td>APPRAISAL</td>
<td>Business &amp; Economics, Archival Science</td>
</tr>
<tr>
<td>319</td>
<td>MODEL</td>
<td>Many disciplines</td>
</tr>
<tr>
<td>269</td>
<td>PERFORMANCE</td>
<td>Business &amp; Economics</td>
</tr>
<tr>
<td>177</td>
<td>BEHAVIOR</td>
<td>Psychology</td>
</tr>
<tr>
<td>137</td>
<td>MANAGEMENT</td>
<td>Management</td>
</tr>
<tr>
<td>125</td>
<td>IMPACT</td>
<td>Environmental Science</td>
</tr>
<tr>
<td>120</td>
<td>INFORMATION</td>
<td>Information Science &amp; Library Science, Computer Science</td>
</tr>
<tr>
<td>113</td>
<td>PERCEPTIONS</td>
<td>Psychology</td>
</tr>
<tr>
<td>105</td>
<td>STRESS</td>
<td>Medical Science, Psychiatry, Psychology</td>
</tr>
<tr>
<td>103</td>
<td>HEALTH</td>
<td>Medical Science, Psychology</td>
</tr>
<tr>
<td>100</td>
<td>RATINGS</td>
<td>Finance, Macroeconomic Management and Sustainable Development</td>
</tr>
<tr>
<td>99</td>
<td>SATISFACTION</td>
<td>Psychology, Management, Medical Science</td>
</tr>
<tr>
<td>88</td>
<td>META ANALYSIS</td>
<td>Medical Science, Psychiatry, Psychology</td>
</tr>
<tr>
<td>82</td>
<td>STRATEGIES</td>
<td>Business Economics, Education &amp; Educational Research, Psychology</td>
</tr>
<tr>
<td>82</td>
<td>RISK</td>
<td>Finance, Medical Science, Business &amp; Economics</td>
</tr>
<tr>
<td>78</td>
<td>WORK</td>
<td>Information Science &amp; Library Science, Archival Science, Psychology, Management</td>
</tr>
<tr>
<td>76</td>
<td>SOCIAL SUPPORT</td>
<td>Psychology, Education &amp; Educational Research, Psychiatry, Medical Science</td>
</tr>
<tr>
<td>76</td>
<td>VALIDITY</td>
<td>Psychology, Psychiatry, Medical Science, Computer Science, Business Economics, Architecture</td>
</tr>
<tr>
<td>73</td>
<td>RESPONSES</td>
<td>Medical Science</td>
</tr>
<tr>
<td>72</td>
<td>SYSTEMS</td>
<td>Computer Science</td>
</tr>
<tr>
<td>72</td>
<td>QUALITY</td>
<td>Education, Environmental Science, Information Science &amp; Library Science</td>
</tr>
<tr>
<td>71</td>
<td>ORGANIZATIONS</td>
<td>Sociology, Management</td>
</tr>
<tr>
<td>70</td>
<td>PERSPECTIVE</td>
<td>Business &amp; Economics, Management</td>
</tr>
<tr>
<td>68</td>
<td>DECISION-MAKING</td>
<td>Business &amp; Economics, Computer Science</td>
</tr>
<tr>
<td>68</td>
<td>FEEDBACK</td>
<td>Computer, Sport Sciences, Education</td>
</tr>
<tr>
<td>67</td>
<td>ACCURACY</td>
<td>Medical Science, Computer</td>
</tr>
<tr>
<td>65</td>
<td>DETERMINANTS</td>
<td>Business &amp; Economics, Medical and Health, Computer Science</td>
</tr>
<tr>
<td>63</td>
<td>CARE</td>
<td>Medical Science, Medical and Health, Hygiene, Library and Information and Digital Library</td>
</tr>
<tr>
<td>61</td>
<td>EMOTION</td>
<td>Psychology, Sport Sciences, Clinical medicine</td>
</tr>
</tbody>
</table>
Figure 7. The mapping of relationship between research focuses from disciplines cited by AAT

Figure 8. The density map of relationships between research focuses from disciplines cited by AAT
Conclusion

The development trends of AAT in the past almost three decades can be extrapolated, and the conclusions can be made, based on the analysis on the disciplines and their research focuses that have cited AAT and been cited by AAT.

Trend 1: the declining influences by Historicism

The emergence and development of archival science was closely related to the study of History. So to speak, archival science in the early stage can be considered as a by-product of History study. The influence of historicism even spurred the formation of “Utilization Determination Theory” of AAT, which regards historians’ actual and expected utilizations of archives as the most important criteria in archival appraisal.

However, the influence by Historicism on AAT has been declining in the past almost thirty year and this trend can be reflected in the table of frequency statistics of disciplines cited by AAT. It shows that the influence on AAT by History study has declined (currently the sixth in the rank of disciplines cited by AAT) and is no longer the most influential discipline to the development of AAT. In the meanwhile, the standards of AAT have become more diversified. Moreover, since the emergence of electronic archives, the attentions paid to electronic archival appraisal exceeded those paid to content (historical paper documents) appraisal. The establishment of “Double Point Theory” for electronic archival appraisal demonstrated the weakening of Historicism in guiding AAT developments. In addition, archives’ properties as technological devices have become popular research focuses of AAT.

Trend 2: the regression to Function Appraisal Theory

Since late 1980s, with the astonishing increase of archival quantities, growing complexity of archival types, and the fast growing applications of archives, archival researchers have been gradually convinced that archives should reflect the societ al values that influenced the archives’ productions. As a result, the ideological basis of archival theory has undergone a fundamental change from being a national theory to being a societal one. However, later on, the focuses of archival appraisal have gradually shifted from an archive’s complex societal values (and the social problems or trends they intent to convey) to the archival provenances shown by the functions of archival creators. The analysis would emphasize on the functions of archival creators, their activities and interrelationships in those activities, and their importance (standings) in business, etc. This new trend, albeit based on the same concept behind the old “Functional Appraisal Theory” (FAT), demands a large body of knowledge from other disciplines to assist archival appraisal, such as Public Administration, Information Storage and Retrieval. The table of frequency statistics of disciplines that have cited AAT and been cited by AAT shows the trend that Public Administration, Information Science & Library Science have been more important than History as far as the influence on AAT is concerned, confirming partially the regressions to FAT.

Trend 3: the predominant popularity of Provenance Principle (PP)

PP is a way to “compartmentalize and manage archives on their provenances to the organizations producing those archives”. It means archival workers arrange and classify the archives according to archival provenance and must stick to the principle that “Do not disperse the archives from the same provenance, and do not mix up the archives from different provenances”. In view of archival value, the principle emphasizes that the archival preservation value should be appraised based on the creators’ functions and the functions of the archives they create. PP had been derived from principle of “Respect General Archives” that originated in France in 1841. The appraisal standards and approaches based on PP are simpler, more objective and practical, hence is an important criterion to judge archival value.

PP made it possible that the applications of AAT to different disciplines. According to the data collected from 1990 to 2016, we could arrive at a conclusion that the disciplines citing AAT and cited by AAT are widely distributed, it shows that all those different disciplines were applying AAT to their own research fields using a variety of different AAT methods, and the methods used depend on the different disciplines and organizations that produce those archives. It reveals that PP is indeed a highly important, or in other words predominant, component of AAT.

Trend 4: the practicalization of appraisal standards

In the past thirty years, the practicalizations of archival appraisal theories has grown to be more
obvious. This trend can be shown in two aspects as follows:

First, the criteria of archival appraisal have been made more favorable to practical application and systematization. In face of the rapid rise of archival quantities, it is very important to make a prompt decision on archival appraisal in time. The procedure for archival appraisal instructions has been made to further detail the storage time of archives in order to facilitate archival appraisal workers to utilize specific methods in a prompt manner. This development trend can be reflected on the research focuses from disciplines that have cited AAT and been cited by AAT, such as the keywords of ORGANIZATIONS, SYSTEMS.

Second, archival appraisal is the main measure for archive optimization. Among a huge collection of files and documents, only a small part of it could be worth being conserved permanently as archives. Therefore, it is essential to control the size of archival collections and to guarantee their qualities at the same time. Archival optimization means to obtain the best proportional relationship between the size of an archival collection and its quantity of information, in other words, how to conserve the maximal amount of necessary information with the minimal size of archival collection. This development trend of archival appraisal can be reflected on the keywords in the articles of collection optimization in AAT, such as MODEL, STRATEGIES, VALIDITY, and QUALITY.

Trend 5: the increasing emphasis on utilitarianism

The keywords of PERFORMANCE and SATISFACTION enjoy very high rankings in the research focuses from disciplines that have cited AAT and been cited by AAT, and it follows that the emphasis on AAT’s utilitarianism (for the utilitarian benefits to the organization that applies AAT) has received more attentions and become an important criterion of archival appraisal.

In addition, evaluating an archive’s value given the cost of conserving it, archival workers try their bests to maximize the benefits applying AAT and minimize the cost. The phenomenon shows that the idea of emphasizing benefits in archival appraisal work has been widely accepted. Archival appraisal is no longer limited to basing itself on subjective demands or on objective properties, but it would also rather be skewed to achieve the optimal benefits in different scenarios and for different disciplines.

Thoughts

The development of AAT is both a dynamic and a regular process. We studied this observation and used the study results to discover the trends of AAT developments in the past almost three decades. These discoveries could be used to augment, tune and enrich Archival Science in order to meet the emerging demands for more applicable and effective theories and methods for appraising modern archival works.

Acknowledgments

The paper is sponsored by the National Social Science Foundation of China (Project No.12BTQ033) and Philosophy and Social Sciences Planning Project of Tianjin City (Project No.TJGL16-011Q and TJTQ16-004).

References

Cook, Terry. (2011). 'We Are What We Keep; We Keep What We Are': Archival Appraisal Past, Present and Future. Journal of the Society of Archivists, 32: 173-189


Elsie T Freeman Finch. (1984). In the eye of the beholder: Archives administration from the user’s point of view. *American Archivist*, 47:112-119


Abstract
This paper proposes a new patent classification method that accounts for both time inflation and application year. We compare the annual citation increment of patents and classify the patents into four types: dominant, precocious, admirable, and late-recognized. Substantially different citation curves were obtained for different patent types. Additionally, early citation history was highly related to the probability of a patent being classified as important.

Keywords
Patent classification, citation analysis, patent analysis

Conference Topic
Patent analysis

Introduction
A patent can be considered a tangible form of an invention or an innovative technology. Among all the indicators employed to measure innovative performance, patent citation and its extensions are the most efficient measurement of patent quality (Breitzman, Thomas, & Cheney, 2002; Costas, van Leeuwen, & van Raan, 2010). When a patent’s technology is based on previously granted patents, these earlier patents are cited in the later patent. If the technology in a previously granted patent is widely used by later inventions, its importance is higher than those that have fewer citations. Thus, a patent’s number of citations indicates its impact or importance.

Long-term forward citations indicate the importance of a patent and its contribution to subsequent research. Younger patents tend to have fewer citations than older patents (Hall, Jaffe, & Trajtenberg, 2001) because of their shorter exposure time. In addition, Hall, Jaffe, & Trajtenberg (2001) proposed that the number of citations a patent receives may be high for reasons other than patent quality; this phenomenon is named citation inflation. When measuring patent quality, different application date and citation inflation may cause citation inequality.

Studies have investigated the distribution of citations during a patent’s life cycle. The shape of the citation distribution is termed the citation age profile, and it can imply knowledge flow, the utilization of knowledge, and the duration of a technology’s life cycle (Mehta, Rysman, & Simcoe, 2010). Patterns of citation durability, aging, and obsolescence vary for different patents (Costas et al., 2010; Glänzel & Schoepflin, 1995; Line, 1993).
Citation inequality caused by time factors should be considered when citations are used to measure patent performance. Most of the current research measuring patent performance has considered only the citations a particular patent receives during the first few years after it was issued (Blind, Cremers, & Mueller, 2009; Breitzman et al., 2002). However, counting citations from only these first few years may overlook the potential impact of patents that are less cited at first but then become highly cited after a long period (van Raan, 2004). Fabry, Ernst, Langholz, & Köster (2006) introduced an impact-factor-like indicator and assigned more weight to citations received in the last two years in their research that was used to position the innovative power and patent strength of companies in the health and nutrition industry.

Various studies have aimed to utilize patent citations in the evaluation of the innovative performance of a single patent, patent assignee, company, or country. However, when examining patent citations as an indicator of patent quality or impact, most studies have failed to account for the time factors of patent citations. Patents applied for in different years may have been examined using the varying policies of intellectual property agencies, and patent application can differ because of the industrial environment during different periods. Comparing patents of different ages is unfair because older patents tend to accumulate more citations. Although there are plausible solutions to managing citation truncation caused by time factors, a comprehensive measurement methodology has not yet been developed.

Numerous studies have employed patent indicators in a diverse analysis; however, the single value represented by patent indicators can only reflect the aggregated performance of a patent until the time at which the analysis is performed. For example, the 2015 accumulated patent citations contain all the citations received by patents up to 2015 if patents of different ages are not distinguished. Therefore, comparing patents of different ages using the 2015 accumulated citations will result in a bias toward older patents and thus an incorrect examination. Moreover, if only static patent indicators are employed—including the accumulated performance during the analysis interval, which does not differentiate the citations gained in each time segment—some interesting behavior of patent may be ignored and the effectiveness of the evaluation will be limited.

In this study, we evaluate patents using their annual patent citation increment, and these values are compared for patents with the same application year. We further propose a classification method for use in future research.

Data
In this preliminary paper, we applied our classification method to patents classified as Information Storage patents by Hall, Jaffe, & Trajtenberg (2005) using the United States Patent Classification (USPC). These USPC included Dynamic Magnetic Information Storage or Retrieval (360), Static Information Storage and Retrieval (365), Dynamic Information Storage or Retrieval (369), and Electrical Computers and Digital Processing Systems: Memory (711). To avoid citation inequality caused by different application years, we compare citations within the same application year. However, in this preliminary research, we only employed 1-year data—grant patents applied for in 1995—to verify our proposed classification method. Citations were counted if a citing patent was applied for from 1995 to 2015 and granted before October 12, 2016, the day we collected our data from the United States Patent and Trademark Office Patent Full-Text Database.
Method
To evaluate the importance of patents during each year, we first divide a patent’s citations by their citation year. For each cited year $y$, we calculate the citation percentile rank and further separate the patents into two groups depending on whether the patents received more or fewer citations than the citation threshold $C_{th,y}$, which is the citation of the chosen percentile in the year $y$.

The following three patent characteristics were considered:
(i) Importance ($g$): the number of times a patent’s annual citations were higher than the citation threshold $C_{th,y}$;
(ii) Stability ($d$): the maximum number of continuous years that a patent received more annual citations than the citation threshold $C_{th,y}$;
(iii) Most influential time ($m$): the age at which a patent received its maximum number of citations.

Figure 1 illustrates the classification method used in this study. If a patent has a significantly higher $g$ than the average (i.e., $g > g_{th}$, where $g_{th}$ is the average of $g$ plus one standard deviation of $g$), we recognize it as an important patent (i.e., dominant). These important patents are separated into two subgroups: stable and unstable. Stable patents are highly cited for a significantly longer period than are others (i.e., $d > d_{th}$, where $d_{th}$ is one standard deviation higher than the average of $d$). For the unstable patents, their citation peak is determined, and they are divided into three subgroups. If their age of maximum citation is significantly larger than those of others (i.e., $m > m_{above}$, where $m_{above}$ is one standard deviation larger than the average of $m$), we term it “late-recognized.” If its citation peak occurs significantly early in its life cycle (i.e., $m < m_{below}$, where $m_{below}$ is the average of $m$ minus one standard deviation of $m$), it is termed “precocious.” An “admirable” patent has $m_{below} < m < m_{above}$.

Result and Discussion
Choosing the Percentile Threshold
To determine the threshold used in the subsequent analysis, we first employed four thresholds (top 1%, 5%, 10%, and 25%) and observed how the classification changed as the threshold was varied (see Table 1 for the results). We discovered that the top 1% threshold was too severe to satisfy and most patents were not recognized as important or stable; this threshold thus resulted in the smallest $g_{th}$ and $d_{th}$. However, we do not expect a dominant patent to remain in the top
1% of cited patents only for a few years. Thus, the top 1% and 5% thresholds could not meet our needs.

### Table 1. Critical indicators for the different thresholds.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>$g_{th}$</th>
<th>$d_{th}$</th>
<th>$m_{above}$</th>
<th>$m_{below}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 1%</td>
<td>1.52</td>
<td>1.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 5%</td>
<td>4.19</td>
<td>3.13</td>
<td>10.99</td>
<td>2.43</td>
</tr>
<tr>
<td>Top 10%</td>
<td>6.74</td>
<td>4.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 25%</td>
<td>11.68</td>
<td>8.33</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Through the critical indicators, the classification results also changed as the threshold varied. As the threshold was gradually changed from top 1% to 25%, more patents were classified as important patents (i.e., dominant, precocious, admirable, or late-recognized; Table 2). However, we noticed that the number of admirable patents increased by a larger amount than did the other important patent types, indicating that a severe threshold will limit the classification result to patents that are extremely important.

### Table 2. Distribution of patent types under different thresholds.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>dominant</th>
<th>precocious</th>
<th>admirable</th>
<th>late-recognized</th>
<th>not-recognized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 1%</td>
<td>118</td>
<td>0</td>
<td>11</td>
<td>15</td>
<td>4115</td>
</tr>
<tr>
<td>Top 5%</td>
<td>222</td>
<td>2</td>
<td>66</td>
<td>52</td>
<td>3917</td>
</tr>
<tr>
<td>Top 10%</td>
<td>359</td>
<td>3</td>
<td>105</td>
<td>63</td>
<td>3729</td>
</tr>
<tr>
<td>Top 25%</td>
<td>509</td>
<td>6</td>
<td>220</td>
<td>76</td>
<td>3448</td>
</tr>
</tbody>
</table>

Figure 2 plots the average citation curve of the four patent types when the four thresholds were used. The gray regions in each plot indicate the citation threshold $C_{th, y}$. When the threshold was top 1%, the annual citations of the citation threshold was higher than even the average of annual citations of dominant patents. Furthermore, we could not separate precocious patents from admirable patents when we applied this threshold. These results indicate that top 1% is not a suitable threshold because it is excessively severe for classifications and can result in the omission or misclassification of some patent types. When the top 5% threshold was used, all four patent types were identified and their characteristic citation curves obtained. However, the citation curves of most of the important patents were still lower than the gray threshold region, and the shapes were strongly affected by the dominant patents. Thus, top 5% is also not a suitable threshold for the classification of these patents. The curves of the top 10% and top 25% thresholds are different from the citation curve for the dominant patents. Both top 10% and top 25% are suitable threshold choices. In the following discussion, we analyze in detail the classification of patents using top 25% as the citation threshold $C_{th, y}$. 
Classification Result

Only a few patents were classified as important patents, as indicated in Table 2. Even if top 25% was used as the threshold, only 19% of patents were recognized as important. However, we noticed that most important patents were stable and classified as dominant. We also discovered that although the citation of most patents peaked approximately five years after application, there were a relatively small number of precocious patents, the citation of which peaked significantly earlier than others.

In Figure 2, the difference in the average citation history of each patent type can be clearly distinguished. The average citation curve of precocious patents peaked early and soon decreased to the average. The dominant patents were not only highly cited for a long period but also gained more citations than other types. The annual citation of dominant patents was two times the annual citation of other important patents on average and more than three times the annual citation of the average citations of all information storage patents. The late-recognized patents have a flat citation curve that is slightly higher than the average and the threshold during the first half of their life; subsequently, the citation curve suddenly increases by a factor of approximately two. After these late-recognized patents were recognized, their citation curve was similar to that of dominant patents but approximately half the amplitude. The average citation curve of admirable patents is similar to the theoretical citation curve of patents, with a peak approximately 5–7 years after the patent’s application.

When the top 25% threshold was employed, most of the average citation curves of important patents (except for the precocious patents during their late life) were higher than the threshold. This indicated that our classification method successfully recognized important patents.

Figure 3 illustrates the classification distribution for four top cited patent groups. For patents having its accumulated citation rank as top 1%, which had a total number of citations larger than 285, we discovered that all were classified as dominant patents. As the accumulated
citation changed from top 1% to 25%, the proportion of dominant patents decreased. Less than half of the top 25% accumulated citation patents were dominant patents. Figure 2 illustrates that the average citation curve varied from type to type. Figure 4 further examines the citation curves of the four types of patent for patents that received the equal number of accumulated citations ($C_{total} = 82$). The patent types in Figure 4 were classified using 25% as the citation threshold $C_{th, y}$. Although these patents had the same number of accumulated citations at the time the data was collected, their citation curves vary distinctly.

Figure 3. Classification distribution under different total citation thresholds.
In the current patent system, assignees must pay maintenance fees to extend the legal status of their patents. Therefore, it is essential to know if a patent is worth maintaining before a fee is paid. If only the total number of citations a patent receives is used to evaluate that patent, some crucial patent information might be missed.

Patent assignees must decide whether to pay the first maintenance fee 7.5 years after the patent was applied for; thus, we attempted to examine the years during which a patent received fewer citations than the citation threshold $C_{th,y}$ during its first 8 years and used our method to classify the patents (Table 3). We discovered that if a patent received fewer citations than the threshold for more than 5 years, it had a small probability (<10%) of being classified as an important patent later in its life cycle. The results presented in Table 3 may help patent assignees to decide whether to pay maintenance fees.

<table>
<thead>
<tr>
<th>below threshold years</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>dominant</td>
<td>89.92%</td>
<td>63.94%</td>
<td>37.38%</td>
<td>21.93%</td>
<td>14.72%</td>
<td>8.52%</td>
<td>4.96%</td>
<td>1.07%</td>
<td>0.27%</td>
</tr>
<tr>
<td>admirable</td>
<td>4.20%</td>
<td>18.27%</td>
<td>26.64%</td>
<td>21.56%</td>
<td>10.74%</td>
<td>5.38%</td>
<td>0.48%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>late-recognized</td>
<td>0.84%</td>
<td>3.37%</td>
<td>7.01%</td>
<td>5.58%</td>
<td>5.52%</td>
<td>3.14%</td>
<td>0.80%</td>
<td>0.00%</td>
<td>0.09%</td>
</tr>
<tr>
<td>precocious</td>
<td>0.00%</td>
<td>0.96%</td>
<td>0.93%</td>
<td>0.00%</td>
<td>0.61%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>
Conclusion

In this paper, we propose a new patent classification method that accounts for time inflation and application year. We compare the annual citation increment of patents applied for in the same year. Thresholds are defined for the identification of significantly highly cited, stable, early cited, and late-cited patents, and they are classified into four types: dominant, precocious, admirable, and late-recognized.

We first investigated the classification results when one of four citation thresholds was employed: top 1%, top 5%, top 10%, and top 25%. We discovered that the top 1% and 5% citation thresholds are too high for most patents to pass. The citation curves obtained when these two thresholds were used were also strongly affected by the dominant patents. Therefore, top 10% or 25% is recommended as the citation threshold when our proposed classification method is employed.

The average citation curves obtained using our classification and the citation curves of selected patents that had the equal number of accumulated citations revealed that the citation curves of different classification types differ significantly. If only accumulated citations are considered, crucial information can be missed.

Our analysis of patent performance in the first 8 years and our classification identified that if a patent has fewer citations than the threshold for more than 5 of its first 8 years, it has a relatively low probability of becoming an important patent. This might help assignees predict the performance of their patent and decide whether to maintain it.

References


https://doi.org/10.1111/1467-9310.00234


https://doi.org/10.1177/016555159502100104


https://doi.org/10.1002/jae.1086
Is the Gap in Scientific and Technological Strength Between G7 and BRICS Becoming Smaller or Larger?

Yuan Junpeng  Gao Jiping  Su Cheng  Zhai Lihua  Wang Haiyan  Pan Yuntao

junpengyuan@gmail.com

Institute of Scientific and Technical Information of China, Beijing (China)

Introduction
Science and technology in emerging economies is undergoing rapid development. As a result, the question arises of whether the gap in scientific and technological strength between emerging economies and developed countries is becoming smaller or larger. Although there are many ways of studying scientific activity systematically, scientometric analysis provides a well-proven means of achieving this goal (Garfield E. 1979). Although numerous lists of highly cited articles have been compiled across a variety of scientific fields, few have included articles from all fields (van Raan 2017, Siddiqi and Usman 2017, Terekhov 2017). The present study was an in-depth scientometric analysis of data. It aimed to identify and analyse the 1000 most highly cited articles and evaluate the research performance of the G7 and BRICS countries. It is hoped that this study will stimulate useful discussion among scientists and research managers about research directions and provide insight for policy and funding decisions.

Dataset and methods
Since the reformation and opening up of its economy in 1978, China has undergone rapid development. The Web of Science (WoS) database was used to compile a citation history of papers for the period 1979–2008. Based on these articles and citations, the top 1000 highest-impact articles in every discipline were selected as the dataset. To measure scientific output, we adopted as indicators the total number of articles, total citation rates, citations per article, and numbers of domestic and international articles. To examine collaboration, we chose as an index of international articles, the international cooperation rate to assess the role of international collaboration in high-impact articles.

Analyses and results
We consider whether a country’s stage of economic development and its S&T level can affect the production of its highly cited papers. Do S&T levels lead to or result from numbers of highly cited papers?

Characteristics of G7 countries’ highly cited papers
Domestic articles are taken to be highly cited papers that were written entirely by authors of one particular country. Analysis of the G7’s domestic papers in 22 disciplines can be used as an indicator of the S&T capability and impact of the G7 countries in various disciplines. It can be seen that domestic papers of the G7 accounted for the majority of the papers in all the disciplines: except for the fields of space science, clinical medicine, agricultural sciences, and mathematics, those proportions were greater than 70%. The proportion of G7 domestic papers to all domestic papers in the discipline of general social sciences was more than 90%

Because it provides access to a wider range of facilities and resources, collaboration is encouraged at a policy level. It enables researchers to participate in networks of cutting-edge and innovative activity. The international cooperation rate can reflect the G7 countries’ status of participation in international scientific and technological cooperation. The international cooperation rate of G7 countries in the 22 disciplines is shown in Figure 1.
than the number of such papers from 1999 to 2008. This is in keeping with the principle that in the G7 countries papers published earlier tend to be cited more often.

Characteristics of BRICS countries’ highly cited papers

In terms of the number of papers in the various disciplines, there was a large gap between the rankings of the five BRICS countries and the G7 countries. Table 1 lists the total number of articles and the ranking by discipline of the BRICS countries.

Table 1. The total number of articles and ranking by discipline of the BRICS countries (1979–2008)

<table>
<thead>
<tr>
<th>Discipline</th>
<th>South Africa</th>
<th>China</th>
<th>India</th>
<th>Brazil</th>
<th>Russia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Materials Science</td>
<td>0</td>
<td>10(7)</td>
<td>3(13)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Geosciences</td>
<td>0</td>
<td>0</td>
<td>1(14)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Molecular Biology &amp; Genetics</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1(13)</td>
<td>0</td>
</tr>
<tr>
<td>Engineering</td>
<td>1(24)</td>
<td>2(17)</td>
<td>3(15)</td>
<td>1(24)</td>
<td>1(24)</td>
</tr>
<tr>
<td>Chemistry</td>
<td>1(21)</td>
<td>0</td>
<td>2(16)</td>
<td>1(23)</td>
<td>1(21)</td>
</tr>
<tr>
<td>Environment/Ecology</td>
<td>2(17)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Computer Science</td>
<td>0</td>
<td>2(18)</td>
<td>3(15)</td>
<td>1(20)</td>
<td>0</td>
</tr>
<tr>
<td>Economics &amp; Business</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Psychiatry/Psychology</td>
<td>1(14)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Space Science</td>
<td>1(15)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Clinical Medicine</td>
<td>0</td>
<td>1(18)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immunology</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Agricultural Sciences</td>
<td>7(16)</td>
<td>1(26)</td>
<td>2(22)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Social Sciences, General</td>
<td>0</td>
<td>0</td>
<td>1(12)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Neuroscience &amp; Behavior</td>
<td>0</td>
<td>1(13)</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Biology &amp; Biochemistry</td>
<td>1(19)</td>
<td>0</td>
<td>2(15)</td>
<td>0</td>
<td>1(19)</td>
</tr>
<tr>
<td>Mathematics</td>
<td>1(25)</td>
<td>3(19)</td>
<td>0</td>
<td>2(21)</td>
<td>2(21)</td>
</tr>
<tr>
<td>Microbiology</td>
<td>1(17)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Physics</td>
<td>1(18)</td>
<td>0</td>
<td>1(18)</td>
<td>1(18)</td>
<td>0</td>
</tr>
<tr>
<td>Pharmacology &amp; Toxicology</td>
<td>2(16)</td>
<td>0</td>
<td>0</td>
<td>1(19)</td>
<td>0</td>
</tr>
<tr>
<td>Plant &amp; Animal Science</td>
<td>4(15)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Multidisciplinary</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The economic strength and S&T capacity of the BRICS countries made substantial progress over the period from 1979 to 2008, and the number of SCI papers of the five countries also increased rapidly. It is evident that the highly cited papers in the BRICS countries were mainly the result of international cooperation. In terms of the average international cooperation rate among the BRICS countries, chemistry was the discipline with the lowest rate; the average rate amounted only to 16.67%. The next lowest were pharmacology & toxicology and mathematics. The other disciplines showed an average rate of greater than 50%. Notable was the rate of 100% in economics & business, immunology, and multidisciplinary, which indicates that the highly cited papers produced by BRICS countries in these three disciplines were entirely the result of international cooperation. Compared with the figures for the G7 countries, the average rate for chemistry and pharmacology & toxicology was also low among the BRICS countries.

Conclusion and discussion

The present study was designed to analyze the published articles that attracted the most citations in the G7 and BRICS countries. Some valuable insights can be gained through this restricted approach.

1) The United States was the world leader, followed by the United Kingdom, Germany, and Japan. There was a significant gap between Italy and the other G7 countries.

2) In terms of the number of papers in the various disciplines, there was a large gap between the rankings of the BRICS countries and the G7 countries.

3) There was a world trend toward domestic research in terms of highly cited papers. The papers in the majority of disciplines in the United States were based on independent domestic research. The United Kingdom, France, Germany, and Italy placed greater emphasis on international cooperation, and Japan and Canada had different tendencies toward international cooperation in different disciplines.

4) The highly cited papers among BRICS countries in economics & business, immunology, and multidisciplinary were produced entirely by international cooperation.

5) The growth rates were high in 13 disciplines in Germany and 12 disciplines in Japan. In the United Kingdom and France, the growth rates were high in eight disciplines. Finally, the United States had high growth rates in five disciplines, and in Canada and Italy growth rates were high in four disciplines.

Acknowledgments

This paper was supported by a grant (No: 71473236) from the National Natural Science Foundation of China(NSFC)

References

Nanotechnology Research Innovation and Commercialisation: Some Salient Aspects

Sujit Bhattacharya¹, Shilpa²

¹sujit@nistads.res.in; sujit.personal@gmail.com
CSIR-National Institute of Science Technology and Development Studies, Academy of Scientific and Innovative Research, New Delhi (India)

²shilpa2796@gmail.com
CSIR-National Institute of Science Technology and Development Studies, Academy of Scientific and Innovative Research, New Delhi (India)

Abstract

Evidence based policy studies provide rationale for developing polices and mechanisms for strengthening research and innovation ecosystem, and for creating strategic roadmaps and new business models. This is becoming critical in science-intensive research areas like nanotechnology where the traditional models of innovation do not work. Taking this as a premise, the paper investigates the contemporary research and innovation trends in nanotechnology. Along with traditional bibliometric indicators, standard activity and product development are examined to provide a deeper insight of nanotechnology research and commercialisation.

Findings confirm with current studies of exponential increase in research papers and aggressive patenting activity in this field. Patenting and linkages between various technology classes/sub-classes and the role of instruments in development of this field shows new insights behind nanotechnology development. Interesting evidence emerges from research activity in production methods which is linked to scalability and commercialisation. Standard making shows how it is a strategic activity with domaination by a few countries. The paper concludes by highlighting the need for examining both macro and micro trends to capture the direction of translation and commercialisation in an new and emerging technology like nanotechnology.

Conference Topic

- Indicators
- Research Fronts and Emerging Issues
- Science policy and research assessment
- Science policy (on different levels)
Introduction

Nanotechnology is seen as the key emerging technology of the 21st century. Radically novel properties that materials demonstrate at the nano-scale (10-9 meter) promises to revolutionize many industry sectors. Nanotechnology has potential for mitigating developmental challenges in water, health, environment among others. Nanotechnology in the last decade or so has attracted global attention due to these promising possibilities leading to articulation of dedicated mission mode programs, roadmaps and other policy instruments by different countries (UNESCO, 2016). Liberal funding had surrounded the various policy initiatives (Battelle, 2013). Advanced OECD countries are major players, however unlike other emerging technologies many newly industrialised and emerging economies have also emerged as influential players. UNESCO (2015) analysis provides good evidence based support the above fact. Research articles in particular demonstrate emerging South economies strong presence. However, patenting activity which is a reflection of a country’s competitive strategic asset shows emerging countries still have a long way to go to match the innovation leaders. As this report shows leading producers of research articles are ranked as China, the USA, India, South Korea, Germany and Japan. However, the order changes when one looks at number of patents per 100 articles. The USA takes the lead with 44, followed by Japan (30), South Korea (27), Germany (22). It is only 2.28 for China, 1.67 for Brazil, 1.61 for India and 0.72 for Russian Federation.

Nanotechnology has made substantial global impact primarily due to the strong public push by different countries (Bhattacharya et al., 2014). Almost no other field has obtained as much public investment in R&D in such a short time as this field. Along with advanced countries, BRICS and other scientifically proficient countries such as Mexico and Taiwan are making huge investments in this field specifically in development of new and improved nanomaterials. Sophisticated instruments have been created that provide advanced tools for engineering nanomaterials.

Nanotechnology at the present stage of commercial development is primarily acting as an enabling technology. Engineered nanomaterials are being embedded to improve the quality of existing products. Different market forecasts for nanotechnology show the market dominance of nanomaterial production. A good estimation of nanomaterials market comes from the second regulatory review on nanomaterials by European Commission (2012). It estimates the nanomaterials market at the global level at around 11 million tonnes, with a market value of roughly 20 Billion Euro. Application domains where nanotechnology is making significant impact are primarily products containing nanomaterials as intermediary products (i.e. functional components such as biosensors or light emitting diodes) or end user products (eg. Nano silver socks or solar cells) whose performance is enhanced by the specific properties of the nanomaterials.

Nanomaterials distinguish itself due to its size which gives it novel properties. However, it is a technical challenge for producing and storing nanomaterials in bulk quantity as the material quickly changes to normal size due to high reactive property a material has at the nano level. Thus, unlike other technologies, nano-scale provides new challenges that make the translation from laboratory to commercialization difficult (Kurva and Mahajan, 2012). The production techniques developed in a laboratory has to be scalable and industrially viable. Distinguishing the issues that emerge from influential studies on nanotechnology innovation and commercialization provides a useful lens for understanding the trends and design effective policy mechanisms.

Assessment of nanotechnology research and innovation based on bibliometrics has been an active area of research. One highly cited article by Roco (2011) gives a long term trend of nanotechnology development. A recent study by Karaulova et al. (2017) shows development
of this technology in Russia and identifies a series of factors that may help to explain Russia’s limited success in leveraging its ambitious national nanotechnology initiative. Hullman (2007), Ghosh and Krishnan (2014), Darvish (2016) among others showed the development of this area and to what extent emerging countries have been able to assert their position in the global stage. Another set of literature available is in analysing relevant set of bibliometrics data in specific application domain like nanobiotechnology Takeda et al. (2009) and clean energy (Wang et al., 2014), graphene (Shapira et al., 2012). Comparing nanotechnology research profile of different countries is visible across various reports and research articles capturing the contemporary trends of that period. Some recent articles that provide contemporary trends include Liu et al. (2017), Gorjiara and Boldock (2017). Some of these indicators include comparing national publication output and impact, international collaboration output and share, contribution and impact of Institutions and impact of journals and patent filings. Some of the scholars have also used sophisticated methods like foresight studies, data-mining and social network analysis to map the structural characteristics and future prospects of this area (for example, Kostoff et al. 2007). Lux research, European commission and various national reports provide important indications of market and commercial trends. Inspite of rich repository of the various studies we find that majority of them have centered on indicators constructed from publications and patents (we call them conventional indicators). The limitation is that they capture more the upstream end of the innovation value chain and thus it calls for revisiting the conventional indicators and also expanding the types of indicators to capture the downstream end of the innovation value chain and the emerging market. The paper attempts to address this limitation by examining research and patent trends that have more relevance to innovation and commercialisation and also examine standards and emerging products in this area.

Methodology

Four proxy indicators namely research publications, patents, standards and products were taken to capture the research, innovation and commercialisation trends in this field. Indicators based on publications provide an important indication of progress in this technology as mastering this technology and developing applications is contingent upon advanced scientific research. In this context, it is important to examine research activity in different nanomaterial and production techniques as they influence downstream activity. Patents help us to judge the inventive ability and a possible indication of developing novel products/processes. Standard making in emerging technologies is a strategic activity as products/processes evolve around a standard. This indicator is generally overlooked inspite of its high significance. Nanotechnology being an enabling technology at this stage, it is difficult to capture the products that have high functional enhancement through nanotechnology. This is an important caveat which should be kept in consideration while capturing nanotechnology based products.

The data for this study was captured through various methods. Publication data was downloaded from Science Citation Index Expanded (SCI-E). Search string defined by Arora et al. (2013) was applied in the title and/or abstract and/or keyword field to capture nanotechnology publications. Thomson Innovation Patent database was used for this extraction of nanotechnology patents. Search string was based on Technology classification provided by the International Patent Classification (IPC) for this field. Standard activity was captured through close reading of various technical documents including International Organization for Standardization (ISO), World Watch Institute (WWI 2006), etc. Woodrow Wilson International Centre for Scholars’ Project on Emerging Nanotechnologies database (www.nanotechproject.org) was used to capture nanotechnology based products in the major
global markets. This database tracks nanotechnology products globally and thus gives a global view of nanotechnology products, industry sectors where nanotechnology is making an impact and also helps to analyse different country’s activity and impact.

Results

Patenting Trends

Patent applications provide an indication of ongoing inventive activity in this field and also an indication of how innovation is taking shape. Patent filing through the Patent Cooperation Treaty (PCT) is the most preferred route for multiple filing in different patent office due to cost effectiveness, ease of multiple filing, establishing and extending priority among others. Nanotechnology patent filing in the PCT is thus a good reflection of global inventive activity in this field. USA being the major market for high technologies has become an obvious choice for global firms to patent in the US patent office for exploiting their invention. Thus a further informed view can be obtained by examining the application trends from US patent office. These considerations led to the examination of nanotechnology patent filing in these two patent offices.

![Figure 1. Year wise Trends in Nanotechnology Patent Applications](chart)

Patent filing shows steady upward growth in the PCT. However, in the US patent office, the shape of patent filing shows erratic changes from 2010 onwards. There would be multiple factors behind this change. Patenting in nanomaterial carbon family particularly graphene and to some extent in CNTs have shown sharp upward trend from 2009 onwards. Also for the first time in 2009 it was observed that investment by industry sector exceed public investment in advanced OECD countries (Battelle, 2012). US market appetite for next generation technologies may have contributed to the sharp upward trend in the USPTO. Can it be that the expected promise of nanotechnology application is not happening and this is having implications for filing patents in the US patent office which is a costly process? Or as the field is consolidating and there is deliberate intervention to focus on commercially and strategically important inventions? Or are their other causal factors behind this decline in patent filing?

It is important to look at the patenting activity in three nanomaterials fullerenes, carbon nanotubes (CNT) and graphene. These three nanomaterials primarily CNT and graphene are most preferred material for enhancing functionality of different products.
CNT provides high strength to weight ratio, electrical resistance and easy penetration through walls which makes it important for applications like light-weight materials, electric cables, solar cells and fabrics. Graphene has similar properties like CNT but shows more promise due to its unique single layer hexagonal lattice configuration. Thus it is increasingly being used in photovoltaics, sensors, optoelectronics, energy storage devices and ultrafiltration devices. Fullerenes acts as a semiconductor, conductor and superconductor under specific conditions and possess ability to form compounds with many different sorts of material with applications in organic photovoltaics, antioxidants, biopharmaceuticals, polymer additives and water purifiers.

Table 1 demonstrates further the technology classes in which patents have been filed in the PCT and the USPTO.

Table 1. Nanotechnology Patent Applications in Different Technology Classes*

<table>
<thead>
<tr>
<th>IPC Code</th>
<th>Description</th>
<th>PCT (USPTO)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2001-14</td>
</tr>
<tr>
<td>B82Y</td>
<td>Specific uses or applications of nano-structures; measurement or analysis of nano-structures; manufacture or treatment of nano-structures</td>
<td>2059 (8320)</td>
</tr>
<tr>
<td>C01B31/02</td>
<td>Preparation of carbon (by using ultra-high pressure, e.g. For the formation of diamonds, B01J 3/06; by crystal growth C30B); Purification</td>
<td>2898 (1894)</td>
</tr>
<tr>
<td>B82B</td>
<td>Nano-structures formed by manipulation of individual atoms, molecules, or limited collections of atoms or molecules as discrete units; manufacture or treatment thereof</td>
<td>3776 (1606)</td>
</tr>
<tr>
<td>A61K9/51</td>
<td>Nano-capsules</td>
<td>1552 (1049)</td>
</tr>
</tbody>
</table>
Patenting activity in different technology classes clearly demonstrate that majority of technology classes cover nanomaterials with increasing activity observed in carbon based nanomaterials. Analysis of the overlapping classes in many patent provides evidence of this. B82B has maximum linkages with C01B31/02 (452 in PCT and 284 in USPTO) showing the active inventive activity in carbon based nanostructures. Also B82Y shows maximum linkages with CO1B31/O2 (170 in PCT and 147 in USPTO) implying the active inventive activity in use or application of carbon based nanostructures.

Table 1 shows that an active area of patenting is in development of scanning probe instruments. Continuous development of sophisticated instruments is pre-requisite for nanotechnology development and novel instruments have high capital value. This plausibly drives invention activity in instruments and creating monopoly through patents. The birth of nanotechnology can itself be traced to the invention of Scanning Tunnelling Microscope (STM) in 1981 (US Patent 4343993) by Gerd K. Binnig and Heinrich Rohrer (Sekhsaria 2013). In 1986, they were awarded the Nobel Prize for this discovery. The STM yields atomic-scale images of metal and semiconductor surfaces which allowed researchers for the first time to probe materials at atomic level. The range of materials that can be imaged with a scanning device increased with the invention of the Atomic Force Microscope (US Patents 4724318 and RE33387) by Gerd K. Binnig in 1986. These two instruments were invented at and with the support of the IBM Corporation, which was interested in scientific advances within the semiconductor industry. Thus industrial research was instrumental for discovery of these two key instruments. The field has progressed with new sophisticated instruments being developed. Also it shows big firms involvement in nanotechnology research. The role of big industry can also be seen in key discovery in this field. This can be seen also in other key discoveries in this field. Carbon nanotubes was discovered by Sumio Tijma in 1991 working at Nippon Electric Company (Shapira, 2012).

Further sub-classification of the technology classes show that substantial inventive activity is observed particularly in the USPTO in “nano-biotechnology or nano-medicine (protein engineering or drug delivery)”. Drug companies are already pushing for nano-enabled interventions for better drug delivery that enhances drug efficacy and has more confined to the affected area. Patenting activity covering quantum dots as markers and nano-optics are also seen. The new imaging technologies and quantum dots enabling new screens as seen in
high value television sets. Thus, patent examination at the level of sub-classes provides insights to the new nano-enabled applications.

Research Trends

Nanotechnology research is getting influenced to a large extent by the downstream activity in the innovation value chain. One of the greatest challenges in nanotechnology commercialization is to produce high quality nano-material, on a large scale at low cost, and in a reproducible manner. Also there is a constant demand for novel nanomaterials. Industry funding and ‘return-to-investment’ demands are directing scientific research towards creating novel nanomaterials that can radically improve the quality of existing products and develop innovative approach in their production process. Also search is for nanomaterials which can transform the next generation technologies.

The exciting possibilities through nanotechnology intervention was initial shown by fullerene, a member of the carbon family. Further research on carbon family led to the discovery of carbon nanotubes and now graphene. Publication trends provide an indication of research activity in the three carbon family nanomaterials (Figure 3). Fullerene was discovered in mid-1980’s after which its publication trend has shown steep growth. Research on fullerene flattened after carbon nanotubes (CNTs) started showing more enhanced properties (Shapira et al., 2012; Izima, 1991). Graphene is now emerging as the most promising nanomaterial because of its unique combination of extraordinary mechanical, electrical, thermal, optical and electronic properties. The steep increase in graphene research is indication of the interest of the research and innovation community. The curve would take a logistic shape in the future because of the possibility of graphene getting more applied as already observed in high-speed electronics and in flexible circuitry.

There was 2473 fold increase in graphene based publications from year 1991-2012; CNTs exhibited 1502 fold increase in publications from 1992-2012; while fullerenes observed only 18 fold increase from 1991-2012. Graphene research is still in nascent stage and evolving very fast. Fullerenes, carbon nanotube, graphene are different forms of nanostructures and changing the form from fullerene to carbon nanotube and graphene has shown the possibility of more promising application. Migration of same research teams from fullerene research to CNT/graphene research is also a major reason behind the changing trend.
Figure 4 highlights countries actively involved in graphene research. Nanotechnology being science–based technologies, research publications (proxy of research activity) is a good indicator of a country’s competency. Figure 4 provides a good indication of how nanotechnology research hot-spots are now dispersed across advanced North as well as emerging South Economies.

![Figure 4a. Top Five Publishing Country in Graphene](image1)

![Figure 4b. Publication output of prolific Institutions](image2)

**Figure 4. Research Trend in Graphene**

Source – Constructed from SCI-E

The difficulties in producing nanomaterials in bulk quantity are to retain nano-size of the particles, proper storage, and limited impurities during the production process. Table 2 gives details of some of the methods extensively used as well as the research trends.

**Table 2. Global Research Activity in Different Methods of Production of Nanomaterials**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Characteristics</th>
<th>Global Research Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2014</td>
</tr>
<tr>
<td><strong>Sol gel</strong></td>
<td>High Chemical Homogeniety; low processing temperatures; controlled size and morphology of nanoparticles; contamination free; varied applications; production rate is very low</td>
<td>548</td>
</tr>
<tr>
<td><strong>Epitaxial Growth</strong></td>
<td>Nanomaterial produced can replace silicon; scalable but expensive; Toxic substances used during production through this method is a major constraint</td>
<td>173</td>
</tr>
<tr>
<td><strong>Exfoliation Method</strong></td>
<td>Large scale production and low cost of production; low yield makes this process not highly suitable for industrial production</td>
<td>134</td>
</tr>
<tr>
<td><strong>Chemical Vapor Deposition</strong></td>
<td>High quality and large scale production of nanomaterials; scalability is low and expensive</td>
<td>69</td>
</tr>
<tr>
<td><strong>Physical vapor Deposition</strong></td>
<td>Flexibility in usage of varied type of organic and inorganic materials for deposition and surface materials;</td>
<td>31</td>
</tr>
</tbody>
</table>
Methods | Characteristics | Global Research Papers | Research Papers  
|---|---|---|---
| Arc Discharge | Large scale production of high quality graphene | 8 | 12 |
| Unzipping of CNT | High quality and low cost for graphene production; impure nanomaterials are produced which is a major constrain | 6 | - |

Source: Research papers extracted from SCI-E; characteristics drawn from various technical bulletins

**Product Development in Nanotechnology**

Table 3 shows how nanomaterials have created impact in different areas. The inventory includes products from 30 different countries. The United States have the most products, with a total of 746, followed by Germany (317), United Kingdom (90), China (58), Japan (56), Denmark (47) and Switzerland (39).

**Table 3. Nanotechnology Based Products**

<table>
<thead>
<tr>
<th>Application Area</th>
<th>Products in each categories</th>
<th>Product Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health and fitness</td>
<td>906</td>
<td>Dryers; Straighteners; Anti-odor Socks; Clothes; Cosmetics; Dressings; Toothpastes; Air Purifiers &amp; Sterilizers; Photo catalysts; Make-up Instruments; Toothbrush; Sunscreens; Tennis Rackets; Leather; Shoes; Glasses; Hair Dye</td>
</tr>
<tr>
<td>Home and Garden</td>
<td>353</td>
<td>Floor Cleaners; Anti-Bacterial Bed Sheets; Paints; Towels; Spot Removers; Luggage; Pillows; Pet Products</td>
</tr>
<tr>
<td>Automotive</td>
<td>211</td>
<td>Watercrafts; Maintenance &amp; Accessories; Paints; Cleaning Products; Dry Waterless Wash; Engine Oils; Lubricants; Metal Sealing</td>
</tr>
<tr>
<td>Cross Cutting</td>
<td>141</td>
<td>Nano-Coatings; Energizers; Anti-fog Products; Antibacterial Kitchenware; Antibacterial – Locks, Makeup Instruments, Shoe Deodorizer, Table ware, Watch, Chain, Taps, Powder; Hair Ion; Lubricants; Computer Hardware’s</td>
</tr>
<tr>
<td>Food and Beverages</td>
<td>117</td>
<td>Micro actives; Aluminum Foil; Antibacterial Kitchenware; Plastic Bottles; Food storage Container; Oils; Supplements; Fruits and Vegetables Cleaner; Filters; Cutting Boards; Cookware’s</td>
</tr>
<tr>
<td>Electronics and Computers</td>
<td>101</td>
<td>Storage/Memory – Chips; Processors; Speakers; Microscope; Hearing Aid; Projection Screen; Cooling Nano Fluid; Germ Free Mouse; Laser Mouse/Keyboards; Waterproof Products; Nano films for Displays and Photonics; OLED’s</td>
</tr>
<tr>
<td>Appliances</td>
<td>65</td>
<td>Active Carbon Silver Nano Filters; Air purifiers; Sterilizers; Steam Iron; Washers; Nano Silver Hair Dryers; Refrigerator; Washing Machine; Microscope; Air Conditioner; Vacuum Cleaner; Lithium-Ion Battery; Nano-in Anti-Fade Detergent; Solar Glass</td>
</tr>
<tr>
<td>Goods for Children</td>
<td>37</td>
<td>Pencils; Baby Carriage; Baby Sunscreens; Plush Toy; Baby Bottle Brush; Baby Mugs; Teeth Developers; Wipes; Baby Blankets</td>
</tr>
</tbody>
</table>

**Standardization**

USA, Canada, Japan, UK and China are seen to be actively involved in standard making in nanotechnology. In the ISO Technical committee for nanotechnology standard, these countries have leading roles as working group chairman of committees/sub-committee.

![Figure 5. TC 229 Set Up and Activities](image)

Source: Constructed from Bhattacharya, 2015

Holding chairmanship of a working group is an indication of different countries acceptance of that country providing lead in development of that standard. For example, China created first standard in nanotechnology which was in material specification in 2005. Subsequently it has developed 21 standards. China has played an important role in the adoption of material specification standard in the TC 229 and its role was acknowledged by the ISO-Technical committee by giving her chairmanship of this group. In International Electro-technical Commission and Institute of Electrical and Electronics engineers (IEEE) also these countries have major role. IEEE has tied up with CMC which includes semiconductors and electronic design automation vendors involved in diffusion of next generation technologies. This strategic collaboration is undertaken to develop *anticipatory standards* for creation of next generation of circuit architectures based on carbon nanotubes and molecular based devices (Bhattacharya 2015). This borrows from their successful model that led to the development of IEEE 802 wired and wireless telecommunications networking standards.

European Union as an entity has significant role in defining standard guidelines applicable to countries belonging to the union. The guidelines provide challenge to non-EU countries to enter their market. U.S. leadership in the international standards-setting process allows it to shape the strategic and technical direction of nanotechnology development everywhere. China’s standardisation activity is component of its overarching strategy for future technology assertion in this critical field. Standard consortia in nanotechnology are forming in these countries and EU. Among other emerging economies we observe nanotechnology standardisation activity in Taiwan, and South Korea are making substantial progress.
Discussion and Conclusion

Research publications provide good indication of research competency. Global research in nanotechnology is addressing cutting edge technology areas (such as sensors, catalyst) as well as areas of developmental challenges (water purification, agriculture, drug delivery, etc). Nanomaterial research is a dominant focus with research shifting in areas of promising cutting edge applications namely carbon nanotube and graphene. Interesting evidence is visible from the publication activity in different production techniques of nanomaterials. Research in increasingly now focussing on new techniques specifically used for production of new nanomaterials like graphene. However, dominating areas of research are still the traditional production methods. Patenting trends do not commensurate with research trends i.e. number of patents filed are significantly less as compared to research publications. However, it is promising to observe strong connect with patenting increasingly happening in areas of developmental concerns. Patents are in nanomaterial application for treatment of various diseases like cancer, nerve disorders, drug delivery systems, fuel cells, solar cells, sensors for biomedical application and environment impact assessment, water treatment and coatings. Nanotechnology applications visible in global market are from different areas i.e. from areas of concern like health and water filters to high technology areas like electronics. Technologically advanced countries are making strong presence in standardisation activity like USA, Japan, South Korea, UK and Canada. China is the only visible developing economy making strong claim in nanotechnology standards particularly covering nonmaterial standards. China’s active involvement in standard creation and adoption in nanotechnology is not surprising as it is a component of its overreaching strategy for future technology domination in the subarea nanomaterials.

Bibliometric and innovation indicators cover one important dimension of development of a technology based area. In future, it will be significant to merge the sophisticated methods with the understanding of these indicators to get detailed information of the structure and dynamics of this area. Some of these techniques could be forecasting and foresight analysis, co-citation analysis, co-authorship analysis and other types of social networks. It will also be interesting to implement micro-level studies covering the development of nanotechnology in a specific region or sector.

References


UNESCO (2016). Nanotechnology is a growing research priority (www.unesco.org)

Patent Citation Inflation: The Phenomenon and its Measurement Methods

Chen Lixin¹ Zhang Lin¹,²* Huang Ying³

¹Dept. Management and Economics, North China University of Water Resources and Electric Power, China
²Centre for R&D Monitoring (ECOOM) and Dept. MSI, KU Leuven, Belgium
³School of Management and Economics, Beijing Institute of Technology, Beijing, China.

Abstract
Recently, the United States Patent and Trademark Office (USPTO) has been granting more and more patents with more and more references. This can lead to serious patent citation inflation. This article undertakes an initial exploration into the phenomenon of patent citation inflation, and presents some concepts and measurement methods. On one hand, patent citation inflation rates were measured using a diachronous method, and the results indicate that citation inflation occurred with a serious impact on the patents granted during 1976–1996. On the other hand, cumulative patent citation inflation rates were measured using a synchronous method, indicating cumulative citation inflation has happened during 1976–2015. Furthermore, the short term citation inflation rates were measured using a synchronous method to detect citation inflation to recent patents granted in recent two years by calculating the rate of change in patent impact factor, similar to journal impact factor, reflecting that citations to the new patents have alternately inflated and deflated during 1979–2015. Taken as a whole, the results of this study verify that patent citation inflation has occurred seriously and frequently over the past four decades, and confirm the ability of established indicators to detect and measure patent citation inflation.

Conference Topic
Patent analysis; indicators; methods and techniques; citation and co-citation analysis

Introduction
Half a century ago, Price (1961; 1963) proposed that the number of scientific publications would grow exponentially. Although the law of exponential growth in science has not proven to be completely correct, traditional scientific research publications are increasing and even new fashioned publications are experiencing very rapid growth (Larsen & von Ins, 2010). Likewise, more and more patents have been granted by the USPTO over the past several decades. In 1976, the USPTO granted 74,966 patents, whereas in 2015, 325,467 patents were granted – a more than four-fold increase. Like scientific articles, patents contain references; applicants are required to appropriately describe and cite prior knowledge about the technological background of their inventions (Criscuolo & Verspagen, 2008; Jaffe, Trajtenberg & Fogarty, 2000). Moreover, during a patent’s prosecution, the examiner must search for any related prior patents and publications, called examiner citations, in order to judge the application’s patentability ( Cotropia, Lemley & Sampat, 2013). Cotropia (2009) and Sampat (2010) suggest that providing an abundance of references could be beneficial for applicants, because examiners may be less inclined to evaluate them effectively and instead presume their validity based on the citations offered. In recent decades, particularly, information retrieval systems and web search engines have been used extensively to help patent applicants prepare comprehensive reference information. It seems the more patents that are granted, the greater the number of references that are listed. In fact, both the total and average number of references listed in each patent increased greatly during the period between 1976 and 2015. In 1976, each successful patent application contained an average of six references, whereas by 2015, the average number of references had increased more than eight times to an average of 50. This suggests that patent references have experienced serious inflation in recent decades.
Patent citation analysis have been widely used to reveal technological knowledge diffusion (Chen & Hicks, 2004; Hu & Jaffe, 2003; Park & Suh, 2013; Nelson, 2009), to trace technological trajectories (Epicoco, 2013; Érdi et al., 2013; Martinelli, 2012), to map technical knowledge domains (Wang, Zhang & Xu, 2011; Weng & Daim, 2012), and to evaluate technology and the capacity for innovation (Albert et al., 1991; Lanjouw & Schankeman, 2004; Verspagen, 2000; Wartburg, Teichert & Rost, 2005). However, when using patent citation analysis methods, the continuing increase in the number of patents has rarely been noted. Moreover, a predilection by applicants to provide a greater number of references within their applications would certainly cause an explosion in patent references and would inevitably result in patent citation inflation. Patent citation inflation may have a significant impact on patent evaluations. For instance, evaluating a patent based on the number of citations it has received, without considering citation inflation in different periods, would not be judicious. Other metrics, such as highly-cited patents and the strength of co-citations, should also be applied cautiously. Hence, in this article, we study the phenomenon of patent citation inflation and present several methods to measure it.

Data

We constructed a bespoke database by downloading USPTO patents granted between 1976 and 2015. Only utility patents, which account for over 90% of all patents granted during the period, were analyzed. Michel and Bettels (2001) found that 90% of the references in USPTO patents were to other USPTO patents. A study by Chen (2017) showed that USPTO patents as references accounted for 61% of the total references provided, foreign patents made up 18%, and non-patent references accounted for 21%. Cotropia et al. (2013) found that the three proportions were 64%, 14%, and 22%, respectively. For simplicity, all references to USPTO utility patents that were granted during 1976–2015 were selected as our dataset. Thus, our dataset contained a total of 5,270,483 utility patents with 59,820,251 references. Table 1 shows the citation counts by year.

Table 1. The number of citations received by patents, grouped by the year the patent was granted.

<table>
<thead>
<tr>
<th>year</th>
<th>issue year</th>
<th>subtotal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1976</td>
<td>869</td>
<td>869</td>
</tr>
<tr>
<td>1977</td>
<td>13,705</td>
<td>14,439</td>
</tr>
<tr>
<td>1978</td>
<td>26,756</td>
<td>40,965</td>
</tr>
<tr>
<td>1979</td>
<td>20,434</td>
<td>47,296</td>
</tr>
<tr>
<td>2012</td>
<td>26,127</td>
<td>4,611,240</td>
</tr>
<tr>
<td>2013</td>
<td>27,962</td>
<td>5,184,186</td>
</tr>
<tr>
<td>2014</td>
<td>28,566</td>
<td>5,638,617</td>
</tr>
<tr>
<td>2015</td>
<td>25,596</td>
<td>5,406,315</td>
</tr>
<tr>
<td>sum</td>
<td>808,420</td>
<td>59,820,251</td>
</tr>
</tbody>
</table>

Methodology

Concepts and formulas

In economics (Abel & Bernanke, 2005; Barro, 1997), inflation means devaluation caused by an increase in the money supply. Inflation leads to a general rise in the price of goods and services and results in a loss of currency value. Patent citation inflation means devaluation in citation counts caused by an increase in the supply of citations. More specifically, it refers to a
rapid increase in the quantity of patent citations, which leads to a rise in the number of citations a patent receives and a subsequent decline in the overall value or impact of those citations. For instance, just because a patent has received significantly more citations than the average for patents issued in other periods does not necessarily mean that the patent is more important or has had more impact, especially if that period is characterized by an unusually high number of citations. A rapid or excessive supply of references – known as patent reference inflation – can result in patent citation inflation which may provide an alternative explanation for the high citation count. Clearly, the year a patent was granted should not play a vital or critical role in whether or not its citation count is considered to be high and, therefore, we have developed several methods to measure patent citation inflation in a given year or period.

Table 2 presents some of basic concepts and computational formulas associated with patent citation and reference inflation. The examples supporting each concept are based on the data in Table 1. Throughout the remainder of the article, in most cases, references refer specifically to a prior USPTO utility patent listed in other USPTO utility patent. The distinction between references and citations is obvious; i.e. a reference is “given to” a patent, whereas a citation is “received by” a patent. In addition, in all the formulas below, \( m \) represents the year under analysis, and \( n \) or \( p \) represents the year of comparison, where \( m \geq p \geq n \geq 1790 \) (the first US patent was granted by President Washington in 1790).

<table>
<thead>
<tr>
<th>Eq.</th>
<th>Description</th>
<th>Formula</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The number of references given to patents granted in year ( n ) by patents granted year ( m )</td>
<td>( R_{m,n} )</td>
<td>( R_{2015,1976} = 25,596 ) the number of times that patents granted in 1976 were referenced by patents granted in 2015</td>
</tr>
<tr>
<td>2</td>
<td>The number of citations received by patents granted in year ( n ) for patents granted in year ( m ); of course, ( R_{m,n} = C_{m,n} )</td>
<td>( C_{m,n} )</td>
<td>( C_{2015,1976} = 25,596 ) the number of times that patents granted in 1976 were cited in 2015</td>
</tr>
<tr>
<td>3</td>
<td>The number of patents granted in year ( m )</td>
<td>( N_m )</td>
<td>( N_{2015} = 299,382 )</td>
</tr>
<tr>
<td>4</td>
<td>The average number of references given to patents granted in year ( n ) by patents granted year ( m )</td>
<td>( AR_{m,n} = R_{m,n}/N_m )</td>
<td>( AR_{2015,1976} = 25,596/299,382 = 0.09 )</td>
</tr>
<tr>
<td>5</td>
<td>The average number of citations received by patents granted in year ( n ) for patents granted in year ( m )</td>
<td>( AC_{m,n} = C_{m,n}/N_n )</td>
<td>( AC_{2015,1976} = 25,596/701,94 = 0.36 ) It is worth noting that ( AC_{m,n} ) is usually greater than ( AR_{m,n} ) since the number of granted patents per year has increased over this forty year period.</td>
</tr>
<tr>
<td>6</td>
<td>The total number of references given to patents issued since year ( n ) by patents granted in year ( m )</td>
<td>( SR_{m,n} = \sum^n R_{m,i} )</td>
<td>( SR_{2015,1976} = 5,406,315 )</td>
</tr>
<tr>
<td>7</td>
<td>The total citations received by patents granted in year ( n ) as of the end of year ( m )</td>
<td>( SC_{m,n} = \sum^n C_{m,i} )</td>
<td>( SC_{2015,1976} = 808,420 )</td>
</tr>
<tr>
<td>8</td>
<td>The average reference rate: the average number of references given to patents granted since year</td>
<td>( ASR_{m,n} = SR_{m,n}/N_m )</td>
<td>( ASR_{2015,1976} = 5,406,315/299,382 )</td>
</tr>
</tbody>
</table>
Equations (6)–(9) show that the cumulative total and cumulative average number of references are not usually equal to the corresponding cumulative citation counts. However, if \( p = n \), they are equal; i.e., \( CSR_{m,n,n} = CSC_{m,n} \) and \( CAR_{m,n,n} = CAC_{m,n} \). For instance, \( CSR_{2015,1976,1976} = 59,820,251 \), and \( CAR_{1977,1976,1976} = 0.11 \).

### Indicators for detecting patent citation inflation

For the purposes of this study, we consider two kinds of inflation: patent reference inflation and patent citation inflation. Patent reference inflation means that the patents issued in a particular period have a higher average reference rate than the patents granted in other periods. The average reference rate is used to determine whether patent reference inflation exists – i.e., if the average number of references listed in the patents that were granted in a year is significantly higher than other years, patent reference inflation is present. For example, in a period from year \( p \) to year \( m \), patent reference inflation can be measured by the rate of change in the average number of references to patents granted since year \( n \) in each patent issued in year \( m \) versus year \( p \). This ratio is calculated by

\[
R_I_{m,p,n} = \frac{ASR_{m,n} - ASR_{p,n}}{ASR_{p,n}} = \frac{\sum_{m}^{p-1} \sum_{i=1}^{n} R_{i,m} - \sum_{i=1}^{p} \sum_{i=1}^{n} R_{i,p}}{\sum_{i=1}^{n} N_{i}}
\]

where \( R_I \) denotes the rate of change of average reference rate. The higher rate of change, the more serious the patent reference inflation. For instance, \( R_I_{2015,2014,1976} = (SR_{2015,1976}/N_{2015})(SR_{2014,1976}/N_{2014}) - 1 = -0.034 \) means that, in 2015, the patent reference inflation rate was \(-3.4\%\) (deflation), caused by the average reference rate declining.

Patent citation inflation means that the patents issued in a particular period were cited more, on average, than patents issued in other periods. Similarly, the average citation rate is the
indicator for patent citation inflation. If the average number of citations received by the patents issued in a given year is significantly higher than other years, patent citation inflation is present. Patent citation inflation for the patents granted in a given year \( m \) can be measured by the rate of change in the average number of citations for each patent issued in year \( n \) and for those in year \( p \) using the following calculation:

\[
CI_{m,p,n} = \frac{ASC_{m,p} - ASC_{m,n}}{ASC_{m,n}} = \frac{\sum_{n}^{m} c_{i,p} - \sum_{n}^{m} c_{i,n}}{\sum_{n}^{m} c_{i,n}}.
\]  

(11)

For instance, \( CI_{2015,1977,1976} = (SC_{2015,1977}/N_{1977})( SC_{2015,1976}/N_{1976}) -1 = 0.043 \) means that, by 2015, the citation inflation rate of patents issued in 1976 compared to those issued in 1977 was 4.3%. In other words, as of 2015, the average citation rate of the patents issued in 1977 was 12.01, whereas that in 1976 was 11.52 – an increase of 4.3%, caused by the average citation rate increasing. Alternatively, we can limit the receiving time of the patents analyzed within a 20-year period after granted in order to avoid bias toward old patents, which have had more time than new patents to garner citations, using the following calculation:

\[
CI_{m,p,20} = \frac{ASC_{m+20,m} - ASC_{p+20,p}}{ASC_{p+20,p}} = \frac{\sum_{m}^{m+20} c_{i,m} - \sum_{n}^{m+20} c_{i,p}}{\sum_{n}^{m+20} c_{i,p}}.
\]  

(12)

For instance, \( CI_{1995,1994,20} = (SC_{1995,1994}/N_{1994})( SC_{1994,1994}/N_{1994}) -1 = 0.083 \), which means the patent citation inflation rate between the patents issued in 1995 and in 1994 is 8.3%; that is, during the period of 20 years after issued, averagely each patent issued in 1995 received 21.89 citations, whereas patent issued in 1994 received 20.21, patent citation inflating 8.3%. The cumulative average of citations/references is another way to explore patent citation/reference inflation. Cumulative patent citations inflation can be measured by the rate of change in the cumulative average of the citations to patents granted since year \( n \) received for each patent issued before the end of year \( m \) and that of year \( p \) with the calculation:

\[
CCI_{m,p,n} = \frac{C_{AC_{m,n}} - C_{AC_{p,n}}}{C_{AC_{p,n}}} = \frac{CAR_{m,n,n} - CAR_{p,n,n}}{CAR_{p,n,n}} = CRI_{m,p,n}.
\]  

(13)

Again, the citations analyzed can be limited to patents that were granted within a 20-year period prior to year \( m \) and \( p \) with the calculation

\[
CCI_{m,p,20} = \frac{C_{AC_{m,n,20}} - C_{AC_{p,20}}}{C_{AC_{p,20}}} = \frac{CAR_{m,m-20,n,20} - CAR_{p,m-20,n,20}}{CAR_{p,m-20,n,20}} = CRI_{m,p,20}
\]  

(14)

The patent cumulative references inflation rate (\( CRI \)) is different from the patent reference inflation rate (\( RI \)). \( RI \) is measured from the view of one year – the year the patents were issued – whereas \( CRI \) is measured over a period, usually many years.

If limiting the citations analyzed to patents that were granted within a 2-year period prior to year \( m \) and \( p \), we can get an indicator to measure short term citation inflation rate, with the calculation

\[
IFI_{m,p} = \frac{(c_{m,m-2} + c_{m,m-2})(N_{p-1} + N_{p-2})}{(c_{p,p-2} + c_{p,p-2})(N_{m-1} + N_{m-2})} - 1.
\]  

(15)

The indicator of short term citation inflation rate (\( IFI \)) actually denotes the rate of change between the patent impact factor, which is similar to journal impact factor, in year \( m \) and in year \( p \), reflecting citation inflation to new patents granted within recent two years.

**Results**

*The explosion of patent citations*

Figure 1 shows the distribution of the total references in patents grouped by the year of issue and the cumulative number of patents granted during 1976–2015, represented as exponents. Figure 1a shows that since 2011 the total number of references (\( SR \)) has increased faster than the cumulative number of patents (\( SN; SR_{p,1976} > SN_{p,1976} \), where \( p \geq 2011 \)). This means that
each patent granted since 1976 averagely has been cited more than once per year since 2011 (see Figure 1b). The USPTO granted 5.27 million patents during the years 1976–2015, and these patents were referenced 5.4 million times in 2015 alone. In recent years, the number of references to patents has increased much faster than the number of patents granted, revealing an explosion in patent citations. Figure 1c shows the distribution of the average reference rate \((ASR)\). Until 2011, the average reference rate had been increasing, but in recent years the rate has remained steady at around 18 \((ASR_{2015,1976} = 18.06)\).

![Figure 1. (a) The distributions of the number of references \((SR_{m,1976})\) and the accumulative number of patents \((SN_{m,1976})\), (b) their ratio, and (c) the average reference rate \((ASR_{m,1976})\), grouped by the year of issue](image)

Figure 2a shows the total number of citations received by the patents granted in year \(n\) \((SC_{2015,n})\). The patents granted in 1976 were cited a total of 808,420 times during 1976–2015. The patents granted in 1998 received the most citations, reaching over 3.35 million. Figure 2b shows the curve representing the average citation rate \((ASC)\). As of 2015, the average citation rate of the patents granted in 1976 is 12; for the patents granted in 1998, it increases to the maximum 23, which is – double that of 1976; for the patents granted in 2015 the rate decreases to nearly zero \((ASC_{2015,2015} =0.04)\), since the patents were rarely cited in the same year.

![Figure 2. (a) The distributions of the sum of citation counts \((SC_{2015,n})\) and (b) the average citation rate \((ASC_{2015,n})\), grouped by the year of issue](image)

If the patents have received significantly more citations, on average, than earlier patents, patent citation inflation is certainly present, especially for the period 1976–1996 in Figure 2b. If patent citation inflation is not present, the patents issued more recently should have received fewer citations, on average, than prior patents, since they have had more time to be cited. The phenomenon that the older patents received fewer citations can be used as an indication of patent citation inflation. Generally, old patents have a higher number of citations than new ones simply because they have had more time to be cited. Hence, an average citation rate curve should decline from left to right. Yet, the curve in Figure 2b is an inverted V-shape. This is a very strange phenomenon, which could be interpreted as patent citation inflation. When a period is characterized by an enormous number of references to other patents, patent citations inevitably increase, and this leads to patent citation inflation. However, according to the
theory of obsolescence in scientific literature (Burton, & Kebler, 1960; Gosnell, 1944; Marton, 1985; Price, 1965), more recent patents have a higher probability of being cited than old patents. Therefore, when patent reference inflation is present, more recent patents should logically have an even higher probability of being cited, but this is not the case. Figure 2b shows that the gap between the peak year (1998) of average citation rate and the latest year (2015) is 17 years.

Figure 3 shows that the curve of cumulative average citation rate (CAC) inclines from left to right over the period 1976–2015, following the average references rate (ASR). In 2015, the average reference rate (ASR<sub>2015, 1976</sub>) reached to 18.06, meaning that each patent granted in 2015 on average supplied 18 references to patent granted since 1976; consequently, by the end of 2015 the cumulative average citation rate had increased to 11.35, meaning that each patent granted since 1976 had received 11 citations averagely. This indicates that the number of references included in patents has been increasing by a significant amount each year. As a result, each patent granted since 1976 has, on average, received an increasing number of citations leading to serious patent citation inflation over the past four decades.

*Figure 3. The distributions of the cumulative average citation rate (CAC<sub>m, 1976</sub>) and the average reference rate (ASR<sub>m, 1976</sub>)*

*Measuring patent citation inflation*

Figure 4 shows the annual patent reference inflation rate (RI<sub>m,m−1, 1976</sub>) during the 20-year period from 1996 to 2015. With the exception of 2012 and 2015, patent reference inflation rates for the period are positive. In 2005, the patent reference inflation rate reached to the maximum 11.52%, i.e., the average reference rate in 2015 was 11.52% higher than in 2004. During the 20-year period from 1996 to 2015, the patent reference had inflated 137.24% (RI<sub>2015, 1996, 1976</sub> = 1.3724; the average reference rate in 1996 was 7.61, whereas that in 2015 reached 18.06 – a more than two-fold increase). This indicates that within the 20-year period patent reference inflation had occurred frequently and seriously – twelve out of the twenty years with inflating more than 5%.

*Figure 4. The distributions of annual patent reference inflation rates RI<sub>m,m−1, 1976</sub> (1996 ≤ m ≤ 2015)*

The patent citation inflation rate CI represents the rate of change in the number of citations over the patents issued in different year. Figure 5a shows the patent citation inflation rates between 1976 and 2015 (CI<sub>m,m−1, 1976</sub>; 1976 < m < 2015). Until 1996, all patent citation inflation rates were positive, whereas the period between 1999 and 2015 shows a decline.
While the citation inflation is clearly indicated between 1976 and 1996, the negative values between 1999 and 2015 do not necessarily indicate citation deflation, because these more recent patents have not had as much time to build their citation count. Figure 5b highlights the period 1976 to 1996 \((CI_{p,p−1,1976};\ 1976 < p < 1996)\), all rates with positive values. Figure 5c shows the patent citation inflation rate \((CI_{p,p−1,20};\ 1976 < p < 1996)\), citations received during 20 years period after issued, all rates with positive values, too. Further, the earlier the patents were granted within the period 1976 to 1996, the less citations on average they have received (see Figures 2, 5b and 5c). It indicates that citations to the patents granted during this 20-year period suffered serious inflation. The curve for \(CI_{p,p−1,1976}\) in Figures 5b is similar to \(CI_{p,p−1,20}\) in 5c. For the indicator \(CI_{p,p−1,20}\), the receiving time of the patents analyzed are limited within a 20-year period after granted in order to avoid bias toward old patents having had more time to garner citations, is more reasonable than the indicator \(CI_{p,p−1,n}\).

Figure 5. The distributions of the patent citation inflation rates \((a)\ CI_{m,m−1,1976};\ (b)\ CI_{p,p−1,1976};\ and\ (c)\ CI_{p,p−1,20}\ (1976 ≤ m ≤ 2015, 1976 < p < 1996)\)

Figure 6a shows the annual cumulative patent citation inflation rate from 1976 to 2015 \((CCI_{m,m−1,1976};\ 1976 ≤ m ≤ 2015)\). All rates are positive, and the curve declines from the left in the early period to the right in the later period. In 2015, the patent cumulative citation inflation rate was 3.7% over 2014. This means that the cumulative average number of citations received by each patent issued since 1976 was 3.7% higher than in 2014. Figure 6b separates the period between 1997 and 2005 \((CCI_{p,p−1,1976};\ 1997 ≤ p ≤ 2005)\), and Figure 6c shows the citations received during the period of 20 years after the patents were granted \((CCI_{p,p−1,20};\ 1997 ≤ p ≤ 2005)\). During 1976–2015, all of the annual patent cumulative citation inflation rates are positive (see Figures 6a), which means that, all patents granted since 1997 have, on average, received an increasing number of citations (see Figure 3).

Figure 6. (a) The distributions of the annual cumulative patent citation inflation rate \(CCI_{m,m−1,1976};\ (b)\ CCI_{p,p−1,1976}\ and\ (c)\ CCI_{p,p−1,20}\ (1976 ≤ m ≤ 2015, 1997 ≤ p ≤ 2005)\)

---

1 Comparing 1976 to 1995 as example, i.e., \(CI_{1995,1976,20} = 2.797\), each patent issued in 1995 averagely received 21.89 citations during the 1995–2015 period of 20 years after issued, whereas that in 1976 received 5.76 during the 1976–1996 period of 20 years after issued, patent citation inflating 279.7% for the two groups of patents, whose issue years with a difference of 20 years. Using Another indicator, \(CI_{2015,1995,1976} = 0.901\), as of 2015, the average citation rate of patents issued in 1976 was 11.52 citations, but that of patents issued in 1995 reached 21.89 – an inflation rate of 90.1%, caused by the average citation rate increasing greatly. Avoiding bias toward old patents having had more time to garner citations, each patent issued in 1976 on average received 5.76 during the 1976–1996 period of 20 years after issued, and \(CI_{1995,1976,20} = 2.797\), means patent citation inflating 279.7%; otherwise, each patent issued in 1976 on average received 11.52 citations during the 1976–2015 period, and \(CI_{2015,1995,1976} = 0.901\), means patent citation inflating 90.1%.
Figure 7 shows the distributions of the annual patent impact factor (PIF) and the annual short term citation inflation rate (IFI), indicating the rate of change in patent impact factor, and reflecting citation inflation to new patents granted within recent two years. The two curves in Figure 7 are similar to each other, and PIF follows and varies with IFI. But IFI varies greatly and are more sensitive than PIF. Distinctly, the change of PIF can indicate the citation inflation to the new patents granted within 2-year period. In 2010, the short term citation inflation rate (IFI) reached 0.268, meaning that the PIF increased from 0.54 in 2009 to 0.68 in 2010 – an increase of 26.8% or inflating 26.8% – for citations to the new patents granted in 2008–2009. In 2011, citations to the new patents granted in 2009–2010 deflated 14.1%. Figure 7 shows that citations to the new patents alternately inflated or deflated, and sometime seriously. But in a long run inflation and deflation counteract each other, so that the citations to new patent have not accumulated to greatly severe inflation during the several decades. In other words, the average number of citations to new patents varies from 0.39 to 0.68, is not suffered serious patent reference inflation in recent decades. It indicates that patent impact factor is a stable indicator in the view of long term in that it can avoid the suffering of patent reference inflation.

Figure 7. The distributions of the annual patent impact factor and the annual short term citation inflation rate $IFI_{m,m-1}$ (1979 ≤ $m$ ≤ 2015)

Interpretation of the Results and the Methods

This article explores the phenomenon of patent reference and citation inflation using a synchronous method and a diachronous method. Patent reference inflation rates (RI), representing the rate of change in the average reference rate, were measured using a synchronous method. The results show that during the period from 1996 to 2015 most patent reference inflation rates reached more than 5%; up until 2011, followed by a sharp decline from 2012 to 2015 in the range of 2.3% to −3.4%. Cumulative patent citation inflation rates (CCI), denoting the rate of change in the cumulative average reference rate, were measured using a synchronous method, indicating cumulative citation inflation has happened during 1976–2015. Cumulative average citation inflation rate is an appropriate, but rough, indicator of the cumulative citation inflation of patents issued over a long period instead of one year. Short term citation inflation rates (IFI) were measured using a synchronous method to detect citation inflation to new patents granted in recent two years by calculating the rate of change in patent impact factor, similar to journal impact factor, reflecting that citations to the new patents have alternately inflated and deflated but in a long run have not accumulated to greatly severe inflation during 1979–2015. This indicates that patent impact factor is a stable indicator in the view of long term in that it can avoid the suffering of patent reference inflation.

Patent citation inflation rates, representing the rate of change in the average citation rate, were measured using a diachronous method. Citation inflation occurred with a serious impact on the patents granted during 1976–1996. All patent citation rates for this period are positive, indicating a year on year increase in the average number of patent citations. However, between 1997 and 2015, all patent citation inflation rates are negative except for 1998. This
means that, year on year, each patent received fewer citations, on average, but this by no means indicates that patent citations suffered deflation. This is a normal, common-sense phenomenon: patents granted more recently have had less time to receive citations and would therefore have fewer average citations.

Unlike synchronous method, diachronous method can only be used to measure citation inflation in patents that have had enough time to accumulate citations. The features and characteristics that accompany obsolescence in scientific literature (Burton, & Kebler, 1960; Gosnell, 1944; Marton, 1985; Price, 1965) mean that a diachronous method does not provide a complete measure of citation inflation rates (CI). When using the diachronous method, a positive patent citation inflation rate (CI) indicates that patent citation inflation is present; otherwise, it cannot indicate that patent citation deflation happens. Factually, patent references suffered serious inflation in the 1997−2011 period, which may lead to more citations for the patents issued during the same period and, consequently, resulting in patent citation inflation. It means that citation inflation rate (CI) is an incomplete indicator with some limitations, which cannot entirely and precisely detect and measure the citation inflation to the new patents granted in recent years especially since patents need time to collect citations when using the diachronous method. However the short term citation inflation rate (IFI), which is synchronous measure, can be used to detect citation inflation to new patents granted within recent two years by calculating the rate of change in patent impact factor.

In addition, according to the theory of obsolescence in scientific literature proposed by many early scholars (Brookes, 1970; Burton, & Kebler, 1960; Gosnell, 1944; Griffith, Servi, Anker & Drott, 1979; Marton, 1985; Price, 1965), scientific literature typically receives many more citations in the first few years of being published. A citation peak is reached several years later, after which gradually fewer and fewer citations are received due to obsolescence. We assert the interaction between patent obsolescence and patent reference inflation is the cause of the strange inverted-V shape in the average citation rate (see Figure 2). When patent references are undergoing inflation, new patents hold the lowest average citation rates because they have not had sufficient time to receive citations. Yet, old patents, that have had enough time to accumulate many citations, see the highest average citation rates, and this is further reinforced by serious patent reference inflation during 1976−1996: the earlier the patents were issued, the less the average citations they received. Patent reference inflation rates, representing the rate of change in the average reference rate, were measured using a synchronous method, with the results showing high patent reference inflation during 1976−2011: until 2011, references to other patents increased greatly each year. Cumulative patent citation inflation rates, denoting the rate of change in the cumulative average reference rate, were also measured using a synchronous method, indicating cumulative citation inflation.

**Conclusions and Discussion**

Over the last forty years, the USPTO has granted an increasing number of patents each year, some of which include a large number of references to other patents. Such an increase in the number of references has led to serious patent citation inflation in this period. This article explores the phenomenon of patent citation and reference inflation using a synchronous method and a diachronous method.

Patent citation inflation rates, representing the rate of change in the average citation rate, were measured using a diachronous method, and the results indicate that citation inflation occurred with a serious impact on the patents granted during 1976−1996: the earlier the patents were issued, the less the average citations they received. Patent reference inflation rates, representing the rate of change in the average reference rate, were measured using a synchronous method, with the results showing high patent reference inflation during 1976−2011: until 2011, references to other patents increased greatly each year. Cumulative patent citation inflation rates, denoting the rate of change in the cumulative average reference rate, were also measured using a synchronous method, indicating cumulative citation inflation.
has happened during 1976–2015. Furthermore, the short term citation inflation rates, denoting the rate of change in the patent impact factor, were measured using a synchronous method to detect citation inflation to new patents granted in recent two years, reflecting that citations to the new patents have alternately inflated and deflated during 1979–2015. Taken as a whole, the results of this study verify that patent citation inflation has occurred seriously and frequently over the past four decades, and confirm the ability of established indicators to detect and measure patent citation inflation.

This article presents some measurement methods to explore into the phenomenon of patent citation inflation. Patent citation inflation rate (diachronous measure), is an incomplete indication of a phenomenon of citation inflation and it can only be used to measure citation inflation to old patents that were issued many years ago because patents need time to accumulate citations. Whereas the short term citation inflation rate, which is synchronous measure, can be used to detect citation inflation to new patents granted within recent two years by calculating the rate of change in patent impact factor. Cumulative patent citation inflation rate (synchronous measure), is an appropriate, but rough, indicator of the cumulative citation inflation of patents issued over a long period instead of one year.

The methods used to measure patent reference and citation inflation in this study are only able to provide a rough indication of patent citation inflation, not a complete and precise picture of the landscape. In future research, we intend to improve our measurements methods to provide a deeper analysis of the causes and effects of patent citation inflation, along with some theoretical methods to eliminate some patent citation inflation’s negative effects.

Acknowledgments
We thank Professor Liming Liang (Henan Normal University) for valuable advices and discussions. This study is supported by the National Natural Science Foundation of China (Grant No. 71573085 and No. 71373252).

References


Larsen, P. O. & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84, 575. doi:10.1007/s11192-010-0202-z.


Abstract
This proposed dissertation research submitted to the Doctoral Forum aims to expand the scope and depth of altmetrics research. On the one hand, the context of web-based digital traces around scholarly works will be explored. On the other hand, this study will expand single platform examinations to show a more holistic picture across multiple platforms and environments. In addition, instead of directly taking counts of occurrences as a proxy for impact, this study will investigate and discuss how the scholarly works have impacted the scholars’ research, teaching, and creative activities in the interpretation of the traces as impact indicators. By incorporating qualitative approaches into metrics research, the pursuit will hopefully contribute to more discussions about the validity and utility of metrics.

Conference Topic
Altmetrics
Science Communication
Participation in science
Twitter

Introduction
Researchers leave traces of their behavior during many stages of their research process. Parts of this process were formerly invisible. Now with scholarship moving online, we can access various types of altmetrics digital traces. Although the term “altmetrics” (Priem, 2010; Priem et al., 2010) is sometimes considered a misnomer (Haustein, Bowman & Costas, 2015) because they are considered complements and not alternatives to citation-based metrics, many theoretical discussion and empirical studies have been conducted in recent years around this concept. Years before the term “altmetrics” was coined, the idea of measuring broader scientific impact on the web had been discussed. The idea of “polymorphous mentioning” in “listservs, electronic discussion fora” and “reading lists or electronic syllabi” is believed to be a critical feature of web-based scholarly communication (Cronin et al., 1998). In addition to scholars being informally cited, Cronin (2005) predicts that “web-based digital objects and usage statistics” will be used to “model scholars’ communication behaviors” and “track their scholarly influence and impact”. In this proposed research, the meaning of the various indicators grouped under the term altmetrics will be discussed from the fundamental concept of digital traces, i.e., the footprints of the scholarly activities.

By investigating footprints of scholarly acts such as of reading, organizing, sharing and discussing scientific works, this study aims to develop a richer picture of scholars and their work. Specifically, I hope to contribute to three main challenges in interpreting the digital traces: 1) to expand current altmetrics beyond simple capture of occurrences of events to include the context in which they occurred, 2) to explore cross-platform behaviors to have a more comprehensive understanding of the research process, and 3) to understand the limitations of the traces as evidence of impact and value.
Related Work

Scholarly Acts
The concept of altmetrics includes a group of metrics based on usage data and social media events related to scientific works. Impactstory aggregates various metrics on the individual researcher level. Their framework of altmetrics (ImpactStory, 2012) includes view, save, discussion, and recommendation. Another similar classification is proposed by Lin & Fenner (2013) for Public Library of Science (PLoS). Haustein, Bowman & Costas (2015) present a framework of “acts” grouped into three categories: accessing, appraising, and applying, with an increasing level of engagement. In this proposed research, the term of “act” is adopted to describe scholarly activities occurring in the context of diverse scholarly communication tools, publishers, online repositories, and general social media. Some of these acts overlap with or influence other acts. For instance, Mohammadi, Thelwall & Kousha (2016) found that 85% of their survey respondents bookmarked articles in Mendeley to cite them in their publications. In a preliminary attempt to enrich the framework of digital traces these acts create, nine groups of 40 acts were defined by Xu & Hemminger (work in progress). Researchers at University of North Carolina at Chapel Hill were invited to rank the cards of scholarly acts in terms of how much they think the acts indicate the scholarly product involved is of value to them. Although the results are yet to be analyzed, this study will inform the further refining of the scholarly acts framework.

Evaluation of Altmetrics
Sud & Thelwall (2014) conclude six methods to evaluate altmetrics: correlation with peer review and other metrics, the sign test, creator motivation interviews or questionnaires, user motivation interviews or questionnaires, source content analysis, and pragmatic evaluations. These methods are a combination of quantitative and qualitative approaches. In terms of the relationship between altmetrics scores and citation measures, positive results have been reported between citations and Wikipedia mentions (Evans & Krauthammer, 2011), Tweets (Eysenbach, 2011; Shuai, Pepe & Bollen, 2012), Mendeley bookmarks (Bar-Ilan et al., 2012; Mohammadi & Thelwall, 2013), as well as other social media including Facebook wall posts, blogs, mainstream media and forums (Thelwall et al., 2013). Most of these studies point out that no matter the correlation coefficients being low or high, the scholarly impact that these two systems of metrics capture are different. As a result, it is necessary to follow up with qualitative approaches to better interpret the impact reflected. Qualitative approaches provide a deeper understanding of not only the scholarly acts but also the confidence of the metrics. For instance, Mohammadi, Thelwall & Kousha (2016) conducted a survey of 860 Mendeley users and found that most (55%) users had read or intended to read at least half of their bookmarked publications, thus concluding that it is reasonable to use Mendeley bookmarking counts as an indication of readership.

Potentially Relevant Theories
In Haustein, Bowman & Costas’s chapter (2015) in a book on theories of informetrics and scholarly communication edited by Sugimoto (2016), they relate the acts to both citation theories and social theories. Citation theories they analyze include normative theory (Kaplan, 1965), social constructivist theories (Gilbert, 1977; White, 2004; Murugesan & Moravcsik, 1978), and concept symbols theory (Small, 1978). Social theories they discuss include social capital theory (Bourdieu, 1985), attention economics theory (Davenport& Beck, 2001), and impression management theory (Goffman, 1959). These theories are potential theories that can be extended using digital traces and metrics to understand value in the scholarly process.
Research Questions

*RQ1: How to interpret the digital traces?*
1. How do technical affordances of platforms and user features affect the interpretation of digital traces?
2. What are the motivations behind the traces?
3. What about other traces that cannot be easily collected (e.g. organizing into personal folders, sharing in e-mails, discussing with a colleague, etc.)?

*RQ2: What does the impact flow look like?*
1. What are the common orders of the acts?
2. Does one act (e.g., tweeting) predict the following act (e.g., citing)? How does one act affect the following act(s)?
3. What is the relationship between the acts and the perceived impact?

Methodology

*Data Collection*
1. Existing documents and artifacts
   a. Altmetric.com data
   b. Twitter data
   c. Mendeley data
2. Online Survey
   a. Targeted audience: The focus of this research will be on some of the most popular acts in terms of the number of captured online events. Considering that Mendeley reader counts (accounting for two-thirds of recent journal articles) and tweets (accounting for one-fifth of recent journal articles) have been shown to be the most prevalent online events captured for scientific papers (Haustein, Bowman & Costas, 2015), this study aims to target the creators of these traces as the target audience of investigation. I plan to firstly retrieve Twitter users who have tweeted at least three scientific articles. Then, those who provide real-world information (name, position, affiliation) in their profile descriptions will be identified and searched to see if they are Mendeley users. At last, I will collect their e-mail addresses on their public home pages, if any. Users who satisfy these three criteria will be put into the pool of targeted survey audience. At last, I hope to get a pool of 5000 randomly sampled users.
   b. Survey logistics: The Qualtrics-based survey will be tested, evaluated and refined through a series of pilot studies. E-mail survey invitations will be sent out to the 5000 potential participants, with a reminder to nonresponding persons in another 2 weeks.
   c. Survey content: The survey will be designed around the following aspects:
      i. Twitter and Mendeley-specific questions. For instance, how many percentages of the saved/tweeted articles are read/cited, do they prefer to send original tweets or retweet, do they usually save articles to organized folders in Mendeley, and so on.
ii. In addition to Mendeley and Twitter, ask them what other commonly (monthly or more frequently) used tools they use in their research, teaching, and creative activities.

iii. Based on their selection of tools, identify their acts and ask them to rank the order of the acts.

iv. Inquire if they are certain of the rationale of their saving/tweeting/blogging/citing scientific articles, to test out if social media based metrics are based on more cursory and perfunctory acts compared to citations.

v. For the types of acts that they can clearly explain the motivations, ask them their common motivations.

d. Semi-structured Interview

e. Interview logistics: At the end of the survey mentioned above, I will ask the participants if they are willing to participate in an additional (compensated) interview and their contact information. The interview will be coordinated via e-mail but conducted via phone. For these interview participants, I will identify and send their tweets of scientific articles and ask them to have their Mendeley library open before the interview. During the interview, I will ask questions based on the tweets and articles saved in their Mendeley.

f. Interview content: The aim of the interview is to gain a more in-depth understanding of the following aspects:

   i. To collect their opinion about these tweeted articles. In other words, have these articles in fact contributed to their research, teaching, and creative activities? If yes, how?

   ii. To gather information about if other types of acts have been conducted to those tweeted articles. For instance, is the tweeted article also saved into personal folders/blogged/discussed in class/shared on e-mail listservs/discussed in a research group meeting/cited?

   iii. To know more about their motivations of all the acts.

Data Analysis

1. Statistical analysis of survey data
2. Markov Model analysis and/or Optimal Matching, and visualization of the sequence
3. Qualitative analysis of interview data

Implications

Theoretical Contribution

This study will explore the story behind the metrics. The impact flow in the research process will be depicted to better explore researchers’ information behavior and motivation. These practices will not only contribute to the development of web-based indicator theory, which is strongly in need, but also inform the revisit of citation theory in this digital era.

Methodological Contribution

This study combines qualitative and quantitative research methods. On the one hand, instead of directly taking counts of traces as a proxy for impact, this study will use survey and interview to collect more information about the traces. How these articles have in fact contributed to their research, teaching, and creative activities will be incorporated in the interpretation of the traces.
as impact indicators. On the other hand, this study will view the traces via a metrics-based lens and extend the work previously conducted on the information behavior of scientists.

**Practical Contribution**

The “altmetrics” data of scientific activities are now being provided in diverse scholarly communication tools, aggregated data providers, as well as publishers and online repositories. Although the data are still far from being well understood, they are being used by individuals, publishers, universities and scientific institutes all around the world in various ways. This study will draw firmer conclusions about the meaning of the metrics, thus providing better guidance on how to use them.

**Doctoral Forum Feedback**

I see four fundamental issues that I would like to discuss at the Doctoral Forum:

- I would like to get any types of suggestions to improve this study.
- I would like to discuss methodological issues of the filtering of survey participants (sampling strategy, technical assistance, and feasibility).
- I would like to get feedback particularly on the design of the survey and interview.
- I would like to brainstorm possible ways to evaluate this study and discuss implications and future directions of this study.

**Statement of Intent**

My name is Shenmeng Xu and I am a doctoral student at the School of Information and Library Science at University of North Carolina at Chapel Hill. My research interests lie in the field of scholarly communication, scientometrics, and altmetrics.

At ISSI 2017, I hope to share my work incorporating qualitative approaches into metrics-related studies, and hopefully contribute to more discussions about the validity and utility of metrics. Although this is not exactly a new perspective, I believe this is an important issue that can never be overemphasized. I hope to bring my expertise in the intersection of scientometrics and scholarly communication to other participants as they work with metrics-related topics in their dissertations.

I would like to attend the doctoral forum because it provides a unique platform to get feedback on my reconceptualization of and approach to understanding the “altmetrics” digital traces. The ISSI community has a rich history of encouraging research that involves studies of scholarly communication and metrics. I believe that the collective of researchers and fellow students in both fields will give me necessary, dual-view insights into the problems I’m grappling with in my work. I plan to write my dissertation proposal this summer and defend it in the Fall semester. An opportunity to attend this doctoral forum will help me strengthen my proposal, and better understand how the ISSI community perceives my research goals.

**Acknowledgement**

I would like to thank my advisor Dr. Bradley M Hemminger for his support and great suggestion on improving this work. I would also like to thank Altmetric.com and its founder Euan Adie for supplying the data and its descriptions.
Reference
Priem, J. (2010). I like the term #articlelevelmetrics, but it fails to imply *diversity* of measures. Lately, I’m liking #altmetrics. 21 September 2010, 3:28 a.m. Tweet.
Abstract
Scientific knowledge, conveyed through the content of scientific literature, is constantly changing as new discoveries and advances are made. These changes could be evolutionary as well as revolutionary changes. Although many population dynamics models and scholarly metrics were successfully applied to characterize the changes of scientific knowledge, accessing scientific knowledge to meet our needs for assessing the state of the art of a research area and making various decisions remains to be a major challenge. In this proposed research, we aim to characterize the scientific dynamics at multiple levels of granularity by analyzing metadata as well as full-text content of scientific literature with various state-of-the-art techniques of graph mining, text mining, and visual analytics.

Introduction
Scientific dynamics conveyed by vast volume of scientific literature is essential for conducting scientific research, making science policies and public understanding of science. The scientific dynamics could be observed at multiple levels of granularity which meet our different needs for scientific knowledge. For example, a scientist and a scientific policy maker may need different information about scientific dynamics for their specific tasks; a researcher who is seeking for research opportunities and a researcher who is conducting a specific research may have different information needs about scientific dynamics.

Today, we still have to build our understanding of scientific dynamics based on scientific literature through painstakingly time-consuming and cognitively demanding processes. The knowledge acquisition process from the vast volume of scientific literature remains to be the most challenging bottleneck not only for scientists and researchers but for everyone who needs to obtain an accurate picture of the state of the art meeting their needs.

The proposed research aims to establish characterize the scientific dynamics at multiple levels of granularity by analyzing metadata as well as full-text content of scientific literature with various state-of-the-art techniques of graph mining, text mining, and visual analytics. The central idea of this research is that structural dynamics of science can be captured by detecting the structure of network constructed by links among scientific literature and that detailed levels (topic, assertion and publication levels) analysis along with their status of uncertainty and novelty in the context of structural dynamics.

Research Objectives
The major objectives of this proposed research are stated as follows:
Constructing evolution model of a domain based on dynamic bibliometric network.

Bibliometric networks, including bibliographic coupling and co-citation networks are widely used approaches to identifying research fronts (Boyack & Klavans, 2010; Shibata, Kajikawa, Takeda, & Matsushima, 2009). For this objective, we construct bibliometric network for each time period and detect communities for each network. To model scientific dynamics over time, we align communities over time. We also extract keywords to characterize detected evolving patterns.

Identifying research concepts and assertions from unstructured text.

The input of this task is a collection of scientific articles in full text. The output is research concepts which constitutes a domain taxonomy and assertions. Assertions in the context of this proposal refer to propositions expressed in the generic form of (Subject)-(Predicate)-(Object), for example, as in (West Nile Virus)-(Causes)-(Persistent Infection). The negation of an assertion is also an assertion. Identified research concepts and assertions will be used to identified detailed level dynamics.

Measuring novelty dynamics of research concepts and uncertainty dynamics of assertions.

Both novelty and uncertainty are early signs and essential features of emerging trends (Rotolo, Hicks, & Martin, 2015). Representing novelty dynamics of research concepts and uncertainty dynamics of assertions fruitfully extends our analytics on science dynamics to a new level of granularity and assist us to keep abreast with the advances of science. Thanks to advances of recurrent neural networks and availability of full-text content, we have chances to represent novelty and uncertainty and measure their dynamics quantitatively.

Integrating analytics techniques for the study of scientific dynamics at multiple level of granularity.

The last objective of my proposed research is to integrate various analytics techniques and display identified information in an explorable interface, which allows users to explore information among multiple levels of granularity according to their needs.

Current Status

In this section, I briefly present three studies which are essential components for my proposed study. The first study presents a visualization for scientific evolution which provide a picture of structural dynamics of science and give a broad context for us to learn detailed level dynamics of science; The second study (work-in-progress) introduces a method to measure novelty dynamics of a research concept; The last study (work-in-progress) introduces a method to represent citation as a vector based on embedding learning.

CiteFlow: Identifying and Visualizing Evolving Patterns in Scientific Literature

CiteFlow is designed to help users track and understand how and why intellectual structure develop and change over time through bibliographic networks of scientific literature. To address this goal, we focus on enabling users to interactively analyse overall network evolution trend, critical events and articles, and text information. Based on these requirements, we designed CiteFlow which consist of four components: a network constructor, a network analyser and an information extractor, and a visualizer.

Figure 1 provides an overview of CiteFlow system. The network constructor is tasked to extract references, abstract, various metadata of research articles, and construct networks. The metadata includes author(s), keywords, publication time and title for each article. The networks include co-citation and co-word networks. The output of the network constructor is dynamic networks with associated metadata and text data. The output is sent to the network analyser.
which extracts a set of communities for network at each time as well as temporal relations of communities (Palla, Barabási, & Vicsek, 2007). Then information extractor extracts the keywords for each detected community and labels them based on abstracts of related articles. Lastly, the visualizer transforms the results from network analyser and information extractor into a comprehensible visualization.

Figure 1. CiteFlow Overview.

Figure 2 shows a visualization case created by CiteFlow. The analysed collection contains articles published from 1966 to 2016 which cited three historical models of science. From this visualization, we can see an evolving patterns of science models. For example, studies on ‘citation network’ emerged in 1980s and got bursty in early 2000s.

Figure 2. Evolution in collection of articles citing three historical models of science (1966-2016): Lotkas law, Goffmans epidemic model, and Prices network model.
Detecting novelty and predict emerging trends in science by word embeddings.

Representing scientific knowledge is a key component of the goal of our proposed research. Notable recent advances of word embedding techniques such as word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and GloVE (Pennington, Socher, & Manning, 2014) have been shown great performance of representing semantics of words. The word embedding techniques are naturally promising to improve our ability to represent scientific knowledge. However, these methods statically represent semantics of each word over all time, making them incapable for interpreting semantic evolution, i.e. scientific knowledge dynamics in our research context. The main challenge of computing temporal word embedding is to represent word semantics derived from different text corporuses across time in a same vector space, which enable us to compare semantic change of research concepts. To resolve this issue, we first constructed embeddings by word2vec in each year and then aligned them by using orthogonal Procrustes over time.

In the next steps, we will measure semantic change (i.e. novelty) of research concepts and evaluate predictive effects of measured novelty on emerging trends. In the future, we may extend word2vec to represent uncertainty between two research concepts and the change of uncertainty over time.

Citation2Vec: Measuring citation similarity based on citation context with embedding learning.

Full-text content of scientific articles leaves us with opportunities to analyse citation context and citation proximity. Citation context is the surrounding text of a citation which describe the author’s interpretation of the cited work; citation proximity is the distance between two citations within the full text of an article which may describe the relatedness of two cited papers. Both of the text information based on citation context and sequential information based citation proximity may be used to describe the cited papers. Inspired by distributed representations of words in natural language processing, we propose a novel approach to represent cited papers by learning from citation context and proximity.

Unlike existed study on distributed representations of scientific articles learning from citation graph or/and text information, our method learns citation representations with citation context and citation proximity. We proposed three methods of learning citation representations from citation context and citation proximity. The ideas of unsupervised learning of citation representations are extensions to word2vec where language model is Skip-Gram model and Negative Sampling is incorporated to optimize loss functions.

The first method is only based on citation context, where embedding vectors for learning are vectors of words surrounding centre citations; The second method is only based on citation proximity, where embedding vectors for learning are vectors of citation vectors surrounding centre citations; The last method is combined first method with second one, where embedding vectors for learning are vectors of words as well as citation surrounding centre citations.

We will compare these three methods based on CITREC (Gipp & Meuschke, 2015), an open evaluation framework for citation-based similarity measurement. CITREC prepares the data of two formerly separate collections for a citation-based analysis and provides the tools necessary for performing evaluations of similarity measures.

Acknowledgments

I acknowledge my doctoral candidacy committee members: Chaomei Chen (Chair, Drexel University, USA), Pak C. Wong (Pacific Northwest National Laboratory, USA), Erjia Yan (Drexel University, USA), and Christopher Yang (Drexel University, USA).

References

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation:


On normalization by division and comparing citation counts

Lawrence Smolinsky*

Department of Mathematics, Louisiana State University, Baton Rouge, LA (USA)

Introduction

Bibliometric measures based on citation counts are often used as an objective measure of research productivity. Finding a method of comparison is extremely challenging given the diversity of scholarly publications particularly by field and age. These are some aspects that effect the way publications accumulate citations. Developing measures using citations depends on understanding the distributions of citations of publications. Methods for normalizing the count across fields for comparison is in constant development.

Division

One heuristically attractive idea is to divide by a version of the expected number of citations to the paper. It may be done for individual papers, institutions, states, etc. Some of the reasons division is attractive is that it is simple to do and transparent in that percentage above or below average may be used in daily life. Some indicators that rely on division are the mean normalized citation score (MNCS) or InCites’s mean category normalized citation impact or normalized citation impact.

Whether division by a mean, median, or other parameter can make citation counts between different areas comparable is dependent upon the distributions being normalized. Each field may be thought to have its own distribution of the number of papers having a given number of citations. These distributions can be interpreted as probability distributions by dividing by the number of papers. For the purpose of cross field comparisons, they should ideally be adjusted so they have the identical distributions for each field. Normalizing by dividing by the mean may give the various distributions the same mean, but otherwise they may be quite different. However, there are families of distributions which become identical with a division adjustment by the mean. In other words, there are well-known families of distributions for which two random variables X and Y in the family with respective distinct means \( m_x \) and \( m_y \) have \( x \sim m_x \) and \( y \sim m_y \) identically distributed. The mean is then a scale parameter. These distributions include the exponential, \( \text{Exp}(\lambda) \), the gamma distribution with fixed \( n \), \( \Gamma(n, \lambda) \), and, a standard candidate in bibliometrics, the lognormal with a fixed \( \sigma \), \( \text{Lognormal}(\mu, \sigma^2) \). Discrete distributions are problematic for scaling since the domain may change. The lognormal scaling was essentially examined in Radicchi, Fortunato, and Castellano (2008) for data in sample areas for the year 1999. They examined the distribution of articles published in a single year in a specific area and approximated it by a lognormal and then compared the various distributions by scaling. The model has its flaws. For example, the lognormal is zero at zero while the number of articles with zero citations is significant. Furthermore, while Radicchi, Fortunato, and Castellano only give an exponential scale graph including multiple areas (2008 Fig. 2 p. 17269), one can visually see some fields below the normalized graph in the tail, e.g., anesthesiology and mathematics. However, suppose one accepts the lognormal with a fixed parameter \( \sigma \) for as the model for the distribution of citations in all fields, i.e., only the parameter \( \mu \) changes from field to field. Then division by the mean for each field normalizes the citation distribution to the common distribution \( \text{Lognormal}(\mu, \sigma^2, \sigma) \) for each field. An indicator like the MNCS, which is a sum of \( n \) normalized random variables (for a fixed \( n \)), will have a common distribution across fields. Even though the distribution may be complicated, it will the same across fields. Furthermore, although the lognormal is sometimes called a thick-tailed distribution, all of the moments of the lognormal exist and the central limit theorem applies to it. Again, assuming that only the parameter \( \mu \) changes from field to field, one can normalize by the median rather than the mean. One has then normalized the citation distribution to the common distribution \( \text{Lognormal}(0, \sigma) \) for each field. We also note that Leydesdorff and Opthof (2011) commented that “.... the ‘mean’ is not a proper statistic for measuring differences among skewed distributions” and suggested using the median. In the lognormal model with a fixed \( \sigma \), both the median and mean are scale parameters.

The discrete lognormal may be a candidate for citations to articles to a single subject and year (Evans, Hopkins, & Kaube 2012; Radicchi, Fortunato, & Castellano, 2008; Thelwall & Wilson, 2014), but it is not clear that \( \sigma \) is fixed across fields. The hooked power is another candidate (Thelwall & Wilson, 2014). But power laws are also candidates (Katz 2016; Yao et al. 2014) and it may depend on the time period. With such variation of opinion and variation of citation distributions, following
principle six of the Leiden Manifesto (Hicks et al. 2015), “Account for variation by field in publication and citation practices” seems a challenge for normalization by division or other formulas.

Percentiles
Knowing the full information of percentiles for two continuous distributions, allows one to transform them into a common distribution. Discrete distributions may pose problems. Eugene Garfield concluded his book (1979 p. 249) commenting on the power of percentile comparisons:

Evaluation studies using citation data must be very sensitive to all divisions, both subtle and gross, between areas of research; and when they are found, the study must properly compensate for disparities in citation potential. This can be done very simply. Instead of directly comparing the citation count of, say, a mathematician against that of a biochemist, both should be ranked with their peers, and the comparison should be made between rankings. Using this method, a mathematician who ranked in the 70 percentile group of mathematicians would have an edge over a biochemist who ranked in the 40 percentile group of biochemists, even if the biochemist’s citation count was higher.

In other words, the proper way to make comparisons would be by percentiles. But this method has its challenges too. It still requires an understanding and integrity of the field distributions. As Garfield (1979 p. 249) continued, “We still know very little about how sociological factors affect citation rates. There is still much uncertainty about all the possible reasons for low citation rates. And there is still much to learn about the variations in citation patterns from field to field.” It is still true today and poses a serious issue if citation analysis is to respect the first principle of the Leiden Manifesto (Hicks et al. 2015), “Quantitative evaluation should support qualitative, expert assessment.” We take the example of one field in which to perform citation analysis and normalization.

Mathematics
Details by subfield classification in mathematics was examined by Smolinsky and Lercher (2012). Even in a single Journal Citation Report field, the distribution of citations in subfields may be very different and at odds with expert assessment. They examined young researchers who were the highest respected by expert assessment, i.e. Sloan fellows. They classified Sloan fellows using a classification by the joint committee of the mathematics, applied mathematics, and statistics societies. The researchers fell into 9 fields (the field of statistics was not included).

In terms of the expert assessment, Algebra seemed to be the highest assessed. It had by far the most Sloan fellows (32 of 99). Hiring of new PhD’s at the top 48 “Group I” mathematics departments was also greatest in algebra. However, this was in contrast to the bibliometric data. The algebra Sloan fellows were 7th out of 9 in the number of publications per fellow and 6th out of 9 for citations per paper.

Putting aside the contradiction between the two assessments, the variation in a single field among subfields is well known. Glänzel and Schubert (2003 p. 357) sum it up: “The classification of scientific literature into appropriate subject fields is...one of the basic preconditions of valid scientometric analyses. Publication activity and citation habits considerably differ among subfields.”

Acknowledgments
The author is grateful to George Cochran for a discussion on scale parameters in probability distributions.

References


Indicators for the Evaluation of Technological Activity

Alexandre Lucas¹, Birger Larsen², Elías Sanz Casado³ & Angel Freddy Godoy Viera⁴

¹alexlucas.al@gmail.com  ²birger@hum.aau.dk  ³elias@bib.uc3m.es
Federal University of Santa Catarina, Florianópolis (Brazil)  Aalborg University Copenhagen, Copenhagen (Denmark)  Carlos III University of Madrid Getafe (Spain)

Introduction

The use of indicators for the evaluation of scientific and technological activities has been widely adopted to characterize and measure its volume and to identify its main actors and investments. One type of approach is to study scientific production which also may also include patents metrics. Another is statistics of governments and international organisation like the OECD that gather information and produce global indicators in their regular reports (González Guitián & Molina Piñeiro, 2008). Despite the large volume of information available, there is a perceived lack of applicability of these data for real time decision making. In this paper, we outline our initial work in establishing dynamic Indicators for the Evaluation of Technological Activity (IETAs) and their potential use in real-time decision making wiht purpose of presenting the idea to ISSI’2017 participants for feedback and exchange of ideas.

Importance

Relying on patent metrics and government data is not sufficient to make real-time decisions about technology investments. The main challenges are:

- Difficulty in obtaining relevant indicators
- Problems with patent-based indicators
- Velocity of technological innovation and the creation of technological companies and products
- The growth of unstructured data
- How to know if some data or information is about technology?
- Open Science and its influence on the future of scientific indicators of economic activities
- Specific issues for the indicators regarding each country - in our particular case for the countries Brazil, Spain and Denmark

Institutes of Science, Technology and Innovation

For our study of IETAs we choose as our case Institutes of Science, Technology and Innovation (ISTIs). ISTIs are important actors within the world of technological development and the innovation ecosystem of each country or even the world. To understand the roles of the involved actors in the phases of development of the technology we use the Technology Readiness Level scale (TRL), initially developed by NASA, but adopted widely by various sectors and organizations including the Horizon 2020 EU program (Mankins, 2009). Fig. 1 shows the TRL scale and our analysis of the types of Research, Development and Innovation (R&D&I) activities and relate the relevant actors across phases. Without discussing the limits of each actor’s position, with the infographic we want to show that the ISTIs are in the middle of the path of technological development. This interval between the TRL 3 and TRL 7 is also known as the “valley of death of innovation” because it is the region with the greatest risks and requirements of investments (Markham et al. 2010).

Fig. 1: Technology Readiness Level and type of R&D and Innovation and related actors in across phases.

About IETA Research

Our research into IETAs started in June 2017 and is planned to be implemented in 4 main steps: 1. Characterization and conceptualization (defining the concept of ‘technology’) and state-of-the-art. 2. Study of Information Retrieval and Text Mining approaches that can support the creation of IETAs. 3. Proposal of principal IETA indicators. 4. Application and test of proposed indicators. In relation to Step 1, we have begun outlining a range of aspects or elements involved in technological activities as shown in Fig. 2. In the present paper the research focus is on information related to technological activity and its flow.
Real-time Indicators

The need for rapid and informed decision-making is increasing fast, and with the current hype around Big Data expectations are building for approaches that can support decision-making in real-time. Fig. 3 shows the lifecycle steps of technology development and the perception of development and impact. As can be seen, it may take long while from an incoming grant or investment until production of the first formal results and end reports.

Expected Results

As indicated in Figs. 2 and 3, we consider a wide range of elements and stages of technological activity. In this paper, we will focus on two examples that involve bibliometric and altmetric data. The first approach is concerned with identifying signs of interest in certain technologies, and we reflect on the possibilities offered in the data related to the Web of Science (WoS). Given a target technology (clearly recognised), expressed as a set of terms as phrases, we may of course search the WoS and analyse publication trends. In addition, the search log of WoS, if made available, can tell us about searcher interest in the target technology (through volume of search terms) as well as which papers are downloaded, e.g. a temporal analysis over the last 5 months. These data may be compared to published papers - allowing indicators and monitoring of early interest to be constructed.

The second approach is concerned with identifying the terms and phrases that characterise of given technology. Initial keyword searches can be conducted in WoS, and fields like Title, Abstract, Keywords, Subject Categories, Journal Names, Project/Grant identifications and Authors and Affiliations can be analysed, e.g. using clustering and visualisation, to extract additional prominent terms to give an overview of the technology aspects and the research fields in which they are relevant as well as provide new search terms for further analysis. In addition, the full text links in WoS to local library content may be utilised to download samples of available full text papers and do in-depth analysis of these using e.g. text mining and clustering. From this, we also hope to provide principal insights into procedures for identifying and characterising technologies - making it easier to so automatically in the future.

Future work and conclusion

Our next step is to carry out the two approaches as outlined in the proceeding section. We also plan to apply information retrieval and text mining techniques to unstructured data from technology and other types of relevant reports. Our aim with presenting these initial project ideas to ISSI’2017 participants is to prompt feedback and to exchange further ideas with the conference participants on how best to proceed.

References


Applying the method of reflections to scientific production in the EU

Wout S. Lamers

Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (The Netherlands)

Introduction
In this paper, we present an initial exploration of the Hidalgo & Hausmann (2009) ‘method of reflections’ applied to scientific production, in particular to European NUTS regions’ scientific production portfolios. This analysis leverages the internal CWTS enhanced Web of Science database to assign publications to geographical areas and classify them into fine-grained scientific topics.

Background
Scientific knowledge production is distributed throughout the world in a highly non-uniform manner. On the whole, the global production of scientific knowledge is concentrated in a relatively limited number of organisations. Similarly, large numbers of journal publications originate from a smaller number of countries, regions or cities. Much like the production of goods, scientific production is highly localized and concentrated, regardless of whether one looks at its geographical or organisational distribution. This suggests that, if we assume that inputs for knowledge production are not distributed nearly as lopsided as its outputs, different organisations, or different regions, must have different capabilities for generating new knowledge.

Method
Hidalgo & Hausmann (2009) show that in economics, product-producer bipartite networks can be analysed to estimate the capabilities of the producers. This so-called ‘method of reflections’ iteratively calculates an average value of the neighbours of a network component. First, diversity of producers’ portfolios and the ubiquity of the produced products are established by counting the edges for each producer and product, respectively. Each reflection subsequently averages the value of a node’s neighbors an increasing number of edges removed from the starting position, resulting in adjusted diversity and ubiquity values.

In this case, we consider European Union NUTS2 regions as producers, and a fine-grained classification of about 4000 scientific topics (see Waltman & Van Eck 2012) as product categories for publications published 2000-2014. Each region is considered to be active in a certain scientific topic if it outputs more publications in this topic than would be expected given its total publication output and the average distribution of publications over the topics. The total number of active topics in a region, and regions in which a topic is active, then constitute the diversity scores for the regions and ubiquity scores for the fields. In this initial exploration, we limit ourselves to the diversity of regions’ scientific production.

Each subsequent ‘reflection’ computes new values – the first step for the regions is the average ubiquity of the regions’ science topics, the second step the average diversity of the regions that share its topics, and so on. This allows us to distinguish between regions with similar sized portfolio of scientific activity; if one region is active in topics that predominantly feature in less diverse regions, and another is active in topics that predominantly feature in diversified regions, the diversity scores in subsequent reflections for the latter region will be higher than those for the former region.

Preliminary results
Figure 1 contains the distribution of the diversity of all 271 regions. London is the most diverse region, followed by Paris. It is clear that a small number of more diverse regions are followed by a continuum of diversity, and the down-sloping tail represents regions with more marginal publication outputs.

A map depicting the regions and the diversity of their scientific production can be found in figure 2. It is clear that the most scientifically active regions are those that contain major centres of population (and, as a result, scientifically active institutions). Beyond this pattern, it seems that the more scientifically active regions are spread across Europe. Most capital regions of Europe show high activity.
Figure 2: diversity rank of EU28 NUTS2 regions

Figure 3 shows the result of the method of reflections, after eight steps through the bi-partite network. This corresponds to a generalized measure of diversification, taking into account the region’s local bipartite neighbours up to 8 steps removed. High values here can be attributed to regions being well-connected to other high-diversity regions and to low-ubiquity topics. It is clear that north-western Europe eclipses south-eastern Europe, and that countries are surprisingly uniform in the ranking of their regions. This may be explained by geographically bound research topics shared by the country’s regions. These outcomes suggest that country-specific capabilities exist, that tie the regions of countries together in the bipartite network.

Figure 3: inferred capability rank of EU28 NUTS2 regions, based on 8th order reflection

Whether these inferred diversity scores of regions correspond to their capabilities may be explored by comparing them to other regional statistics. We check if regions’ diversity, and diversity-related reflection outcomes, correlate to regional GDP per inhabitant, and to regional purchasing power. See table 1 for the outcomes.

Table 1: correlation between measures of diversity and regional purchasing power and GDP per head of population

<table>
<thead>
<tr>
<th></th>
<th>R0</th>
<th>R2</th>
<th>R4</th>
<th>R6</th>
<th>R8</th>
<th>R10</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP/p</td>
<td>0.41</td>
<td>0.47</td>
<td>0.54</td>
<td>0.57</td>
<td>0.56</td>
<td>0.52</td>
</tr>
<tr>
<td>GDP/p</td>
<td>0.39</td>
<td>0.48</td>
<td>0.59</td>
<td>0.66</td>
<td>0.69</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Table 1 shows increasing correlation, to a point, for both regional GDP per inhabitant and purchasing power, as the method of reflections progresses. While this is of course not a definitive confirmation of our capability-related view on regional knowledge production, it does suggest that knowledge production networks can be mined for additional information on the knowledge producers and how their portfolio’s relate to one another. While regions make an attractive visual case for demonstration purposes, universities and other research organisations might be better suited for this kind of analysis as their ‘capabilities’ might be easier to quantify and more readily influenced by institution (hiring and funding allocation) policy.

References


How many keywords do authors assign to research articles – a multidisciplinary analysis?

Jin Mao¹  Kun Lu²  Wanying Zhao³  Yujie Cao³

¹danveno@163.com
School of Information, University of Arizona, Harvill Building, 1103 2nd St., Tucson, AZ, 85719, USA

²kunlu@ou.edu
School of Library and Information Studies, University of Oklahoma, 401 W. Brooks St., Norman, OK, 73019, USA

³wanying-chiu@foxmail.com, cathy0021@163.com
Centre of Information Management, Wuhan University, No.299 Bayi Street, 430072, Wuhan, China

Introduction
Co-word analysis has been widely used to reveal knowledge structures of a research field (Callon et al., 1983). Semantic items used in the co-word analysis are assumed to represent the content of articles, such as terms from titles (Bhattacharya & Basu, 1998) and abstracts, or subject terms assigned by professional indexers (Ocholla et al., 2010). Author keywords have also been selected as one crucial data source for establishing co-occurrence relationships between keywords in many co-word analysis studies (Cho, 2014). The number of author keywords influences the results of co-word analysis. In the co-word analysis, not all author keywords are used to form co-occurrence relations. Only the keywords from papers that have two or more keywords are used. Therefore, investigating on the number of author keywords in research articles will help know the coverage of papers by the co-word network using author keywords. Author keywords are often requested or encouraged by many journals to highlight the content of articles in their publication policies. Some journals even specify the number of keywords author should assign. However, in a research field, what is the distribution of author keywords in papers? Further, to our best knowledge, there are no studies that compare author keywords distributions in multiple disciplines. In this study, we will investigate on how many author keywords are assigned to articles in multiple disciplines with the aim of revealing the underlying quantitative patterns.

Data and methods
Six research fields from soft science to hard science were analysed, including Ethnology, Sociology, Library and Information Science (LIS), Economics, Physics, Fluids & Plasma (Physics), and Acoustics. Articles from top journals in these six fields were retrieved from the Web of Science database. Bibliographic records of articles with the types of article were downloaded. Author keywords in the papers were parsed and stored into MySQL databases. The number of keywords for each article in the six fields was counted. Normal distribution, Poisson distribution, and Weibull distribution were fitted for the number of keywords per paper by applying Maximum Likelihood Estimation. Chi-Square(χ²) tests and Kolmogorov-Smirnov(KS) tests were applied to evaluating the goodness of fit.

Results and discussion
The six disciplines have different magnitudes of selected research articles (Table 1). Table 1 shows that a significant proportion of papers across all these fields do not have any author keywords. Sociology has the lowest rate of keyword absence with 22.56% and Acoustics has the highest rate with 73.84%. In addition, very few papers in these fields have only one keyword. Generally speaking, research fields in hard science (e.g., Physics and Acoustics) tend to have more papers without author keywords than those of soft science (e.g., Ethnology and Sociology).

The significant portion of papers without any keywords or with one keyword in the six fields indicates that only using author keywords in co-word

<table>
<thead>
<tr>
<th>Disciplines</th>
<th># of papers</th>
<th># of papers without keywords/percentage</th>
<th># of papers with one keyword/percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnology</td>
<td>6,655</td>
<td>2,211 (33.22%)</td>
<td>3 (0.05%)</td>
</tr>
<tr>
<td>Sociology</td>
<td>7,307</td>
<td>1,719 (23.53%)</td>
<td>2 (0.03%)</td>
</tr>
<tr>
<td>LIS</td>
<td>12,010</td>
<td>4,200 (34.97%)</td>
<td>53 (0.44%)</td>
</tr>
<tr>
<td>Economics</td>
<td>71,455</td>
<td>47,492 (66.46%)</td>
<td>104 (0.15%)</td>
</tr>
<tr>
<td>Physics</td>
<td>63,155</td>
<td>43,168 (68.35%)</td>
<td>184 (0.29%)</td>
</tr>
<tr>
<td>Acoustics</td>
<td>93,451</td>
<td>68,011 (72.78%)</td>
<td>61 (0.07%)</td>
</tr>
</tbody>
</table>
analysis could suffer from the short of paper coverage in the fields. One major reason for the absence of keywords could be the publication rules of the journals. Many journals do not allow authors to assign any keywords to articles, such as *Acta Acustica united with Acustica* in Acoustics and *The World Economy* in Economics. Some journals did not allow keywords in earlier years but started to require author keywords more recently. For example, *Applied Acoustics* did not ask for author keywords before 1995, but later did. Another factor comes from the authors. Even some journals require author keywords, some papers still have no keywords. For example, 14 of 40 papers published in *Youth & Society* in 2014 have no keywords.

Figure 1 presents the author keyword distributions in the six fields. The number of papers is logarithmized.

![Figure 1. Empirical author keyword distributions in the six fields. The number of papers is logarithmized.](image)

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Poisson</th>
<th>Normal</th>
<th>Weibull</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnology</td>
<td>0.2677</td>
<td>0.2441</td>
<td>0.2326</td>
</tr>
<tr>
<td>Sociology</td>
<td>0.3208</td>
<td>0.2069</td>
<td>0.2005</td>
</tr>
<tr>
<td>Economics</td>
<td>0.3136</td>
<td>0.1966</td>
<td>0.1777</td>
</tr>
<tr>
<td>LIS</td>
<td>0.2662</td>
<td>0.1747</td>
<td>0.1601</td>
</tr>
<tr>
<td>Physics</td>
<td>0.2832</td>
<td>0.1695</td>
<td>0.1626</td>
</tr>
<tr>
<td>Acoustics</td>
<td>0.2929</td>
<td>0.1974</td>
<td>0.1891</td>
</tr>
</tbody>
</table>

*Bold values indicate the smallest distances.*

Table 2. The distance values by KS tests

Conclusions

Author keywords in six fields were analysed quantitatively. Results show that a large portion of papers have no keyword or only one keyword. The six fields share similar author keyword distributions with Weibull distributions best fitted although not strictly. The quantitative patterns of author keywords provide practical implications to keyword selection in co-word analysis and reveal the underlying mechanisms in scientific communication. In future, journal rules on author keywords will be further analysed to interpret the patterns. In addition, we will explore whether and how co-word analysis is influenced by the patterns of author keywords.

References


Investment and results in R&D in the temporal context of the economic crisis: a metric approach

Carolina Callejo-Lavado¹ Zaida Chinchilla-Rodríguez² Benjamín Vargas-Quesada³

¹ carolinacallejolavado@gmail.com
University of Alcalá (UAH), Libreros, 21, 28801 Alcalá de Henares (Madrid), Spain

² zaida.chinchilla@csic.es
CSIC, Institute of Public Goods and Policies (IPP), SCImago Research Group, Albasanz 26-28, 28037 Madrid, Spain

³ benjamin@ugr.es
University of Granada (UGR), SCImago Research Group, Campus Universitario de Cartuja 18071 Granada, Spain

Introduction
With the beginning of the global economic downturn in 2008 many countries have reduced their investment in R&D. The tendency of nations’ scientific spending has traditionally been procyclical, being positively related to the level of economic activity (OECD, 2016). Although the causal relationship is not guaranteed (Leydesdorff & Wagner, 2009), there is a lot of scientific literature that positions R&D investment as a driver of productivity and long-term economic growth in the industrially advanced nations (García-Porras et al., 2015). Although higher investment does not always guarantee better results (King, 2004), because scientific policy plays a key role, insufficient funding can lead to research deficiencies (Reed et al., 2007).

The objective of the present work is to explore the effect that cuts in the R&D investment may be having on the results of scientific production and its performance at the international level.

Methods
An exploratory analysis is carried out on inputs-outputs. The countries studied have been selected according to their level of expenditure in R&D: Japan, Finland and Sweden (high R&D effort); Slovenia and Czech Republic (increasing effort); Spain, Italy and Poland (low effort); France, Netherlands and United Kingdom (average effort). The period of analysis covers the pre-crisis stage (2000-2007) and the post-crisis stage (2008-2015). As a measure of the variables, the Average Annual Growth Rate (AAGR) is used by time series. For the input, the study variable is the expenditure on R&D (% of GDP), with data extracted from the OECD. For the output, some bibliometric indicators that quantify the performance of scientific production: Normalized Impact Index (NI), percentage of Leadership (L), percentage of Excellence (Exc.) and percentage of Excellence with Leadership (EwL), all extracted from Scopus, through the access support of SCImago Institutions Rankings (SIR).

Results

Inputs (R&D expenditures)
Finland, United Kingdom, Spain and Japan performed, on average, a greater economic effort in the years before the crisis and reduced it in the later years (Figure 1). United Kingdom highlights, with a negative growth in both periods. Italy is the only nation whose AAGR has been similar before and after the crisis. The remaining countries (Sweden, France, Netherlands, Czech Republic, Poland and Slovenia) have intensified their investment in the years of economic recession compared to the previous period, some significantly (the last three).

Figure 1. AAGR of GERD as percentage of GDP in the pre-crisis and the post-crisis period

The analysis by time series (Figure 2) shows that Finland, Japan, Poland, Spain and United Kingdom made an increasing R&D effort during the first years of the decade, but from 2009 there was a coincidental decrease with the years of economic instability. In Slovenia the same has happened but from 2012. Czech Republic and France have had...
similar behavior, reducing the growth rate of their investment from 2009 after a few years of increase. Sweden and Italy have been changing, with some periods of increase and others of decrease in their growth; while in the Netherlands spending was negative before the crisis, but positive from 2009.

Figure 2. AAGR of GERD as percentage of GDP by time series

Outputs (R&D results)

Analysis of scientific results by time series shows that R&D cuts, caused by the economic downturn, may be negatively affecting the growth of leadership (Figure 3) and excellence with leadership indicators (Figure 4). In all the countries analyzed, percentage of leadership has decreased over time; although it is true that this has been more pronounced since 2009 in almost all countries (Finland, Sweden, Czech Republic, Slovenia, Poland, Spain, France, the Netherlands and United Kingdom). This has affected the rate of excellence with leadership, which has also decreased especially during the crisis in Finland, Japan, Sweden, Spain, France, Netherlands and United Kingdom, so some countries are losing scientific autonomy and rely on others to produce high-quality research.

Figure 3. AAGR of Normalized Impact and % Leadership by time series

The growth trend of excellence (Figure 4) has been predominantly positive in some countries (Czech Republic, Slovenia, Italy and Poland) and even increasing in the years following the recession. Japan and United Kingdom went from a negative growth trend to a positive trend in the latest time series. However, in other countries the growth rhythm has slowed since 2009 (Spain and Italy), even becoming negative from 2012 (Sweden, France and Netherlands).

Figure 4. AAGR of % Excellence and % Excellence with Leadership by time series

The indicator whose growth has been least affected by cuts in R&D is the normalized impact index (Figure 3). In some countries it has remained more or less stable with positive rates (Japan, Sweden, Italy, Spain and United Kingdom) and in others it has grown more in the crisis years than in the previous years (Finland, Czech Republic, Slovenia and Poland). France and Netherlands are the only ones with a tendency to decrease in their impact from 2009. Further research should be done to better understand the potential effects on funding and research performance in order to foster the best political and academic strategies that might help for building the research agendas of nations.

References


Reed, D. A.; Cook, D. A.; Beckman, T. J., et al. (2007). Association Between Funding and Quality of Published Medical Education Research. JAMA, 298, 1002-1009.

1 Time series may vary due to data availability.