



## King's Research Portal

DOI:

[10.1109/TCOMM.2023.3255252](https://doi.org/10.1109/TCOMM.2023.3255252)

*Document Version*

Peer reviewed version

[Link to publication record in King's Research Portal](#)

*Citation for published version (APA):*

Du, W., Tan, T., Zhang, H., Cao, X., Yan, G., & Simeone, O. (2023). Network Topology Inference Based on Timing Meta-Data. *IEEE Transactions on Communications*, 71(6), 3263-3273.  
<https://doi.org/10.1109/TCOMM.2023.3255252>

### **Citing this paper**

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

### **General rights**

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

### **Take down policy**

If you believe that this document breaches copyright please contact [librarypure@kcl.ac.uk](mailto:librarypure@kcl.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.

# Network Topology Inference Based on Timing Meta-Data

Wenbo Du, *Member, IEEE*, Tao Tan, Haijun Zhang, *Senior Member, IEEE*,  
Xianbin Cao, *Senior Member, IEEE*, Gang Yan, *Member, IEEE*,  
and Osvaldo Simeone, *Fellow, IEEE*

## Abstract

Consider a processor having access only to meta-data consisting of the timings of data packets and acknowledgment (ACK) packets from all nodes in a network. The meta-data report the source node of each packet, but not the destination nodes or the contents of the packets. The goal of the processor is to infer the network topology based solely on such information. Prior work leveraged causality metrics to identify which links are active. If the data timings and ACK timings of two nodes – say node 1 and node 2, respectively – are causally related, this may be taken as evidence that node 1 is communicating to node 2 (which sends back ACK packets to node 1). This paper starts with the observation that packet losses can weaken the causality relationship between data and ACK timing streams. To obviate this problem, a new Expectation Maximization (EM)-based algorithm is introduced – EM-causality discovery algorithm (EM-CDA) – which treats packet losses as latent variables. EM-CDA iterates between the estimation

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFF0301400, in part by the National Natural Science Foundation of China under Grant 61961146005, in part by the Shuohuang Railway Project under Grant GJNY-19-90. The work of O. Simeone was supported by the European Research Council (ERC) under the European Union’s Horizon 2020 Research and Innovation Programme (Grant Agreement No. 725731) and by an EPSRC Open Fellowship.

W. Du, T. Tan and X. Cao are with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China, with the Key Laboratory of Advanced Technology of Near Space Information System (Beihang University). (e-mail: wenbodu@buaa.edu.cn; tantao@buaa.edu.cn; xbcao@buaa.edu.cn).

H. Zhang is with Beijing Engineering and Technology Research Center for Convergence Networks and Ubiquitous Services, University of Science and Technology Beijing, Beijing, China, 100083 (e-mail: haijunzhang@ieee.org).

G. Yan is with School of Physics Science and Engineering, Tongji University, Shanghai 200092, China (e-mail: ee-yan@gmail.com).

O. Simeone is with the King’s Communications, Learning, and Information Processing (KCLIP) Laboratory, Department of Engineering, King’s College London, London WC2R 2LS, U.K. (e-mail: osvaldo.simeone@kcl.ac.uk).

of packet losses and the evaluation of causality metrics. The method is validated through extensive experiments in wireless sensor networks on the NS-3 simulation platform.

### Index Terms

Network topology inference, meta-data, causality metrics, packet loss, expectation maximization.

## I. INTRODUCTION

### A. Motivation and Overview

Information about the topology of a device-to-device wireless network, e.g., a sensor network, is essential to implement functionalities such as routing, anomaly detection, and load balance. In recent years, *passive* monitoring methods that leverage only observations of network traffic have received significant attention, owing to their cost-effectiveness as compared to *active* methods that probe nodes for information [1], [2]. Passive monitoring methods can be “invasive”, implementing packet inspection techniques like demodulation and decryption [3]; or “non-invasive”, leveraging only meta-data. Invasive methods can achieve high accuracy, but they require complex sensors and baseband processors. Non-invasive techniques have the advantage of requiring only information about the timings of data packet and acknowledgement (ACK) packets, which is relatively easier to collect and process (see Fig. 1). This paper contributes to the line of work on passive, non-invasive, network topology estimation.

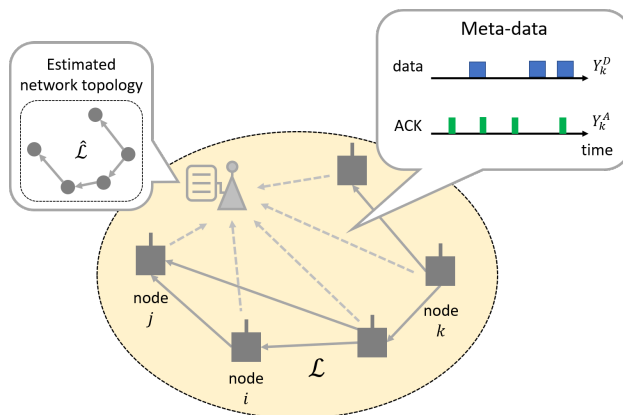


Fig. 1. An example of a wireless device-to-device network with a set of nodes  $\mathcal{N} = \{1, 2, 3, 4, 5\}$  and a set of directional links  $\mathcal{L} = \{(2, 1), (3, 1), (3, 2), (4, 3), (4, 5)\}$ . The central monitor collects meta-data from the nodes in the form of data and acknowledgment (ACK) packet timings, based on which it aims to estimate the network topology. Unlike previous work [4], [5], this paper allows for packet losses, making it more challenging to interpret and use meta-data.

To elaborate, consider, as in Fig. 1, a processor having access only to meta-data consisting of the timings of data packets and ACK packets from all nodes in a network. The meta-data report the source node of each packet, but not the destination nodes or the contents of the packets. The goal of the processor is to infer the network topology based solely on such information. Reference [6] proposed to leverage causality metrics to identify which links are active. The key underlying idea is that, if the data timings and ACK timings of two nodes – say node 1 and node 2, respectively – are causally related, this may be taken as evidence that node 1 is communicating to node 2 (which sends back ACK packets to node 1). The same principle underpins network discovery in fields as diverse as biology and sociology [7]–[10].

The causality discovery algorithm (CDA) introduced in [6] was based on Granger causality, a measure of causal dependence based on auto-regressive modelling [11]. Asymmetric Granger causality was used in [4], which outperforms GCT at a finer time resolution. Transfer Entropy (TE) was then adopted for CDA in [5]. TE has the advantage of capturing also non-linear causality relationships [12]–[14].

This paper starts with the observation that packet losses can weaken the causality relationship between data and ACK timing streams. To obviate this problem, a new Expectation Maximization (EM)-based algorithm is introduced – EM-causality discovery algorithm (EM-CDA) – which treats packet losses as latent variables.

## *B. Related Work*

Active probing is a traditional method used in wireless topology inference, whereby information is collected from neighboring nodes [15], [16]. In such methods, a subset of “privileged” nodes usually performs the probing task [17]. While these schemes can potentially infer accurately the functional network without location information, the energy cost associated with active methods is a critical drawback.

As for passive schemes, references [18], [19] exploit spectral coherence to infer the network topology, but this approach tends to detect spurious links. In [20], [21], multivariate Hawkes processes, a parametric formulation of packet arrival statistics, is considered to recover the network topology. These solutions are model-based, and hence operate under strict assumptions on the valid of the model. CDA-based passive topology inference methods currently provide state-of-the-art results for passive topology inference. Apart from the papers reviewed in the previous subsection, the authors of [22] leverage blind source separation to improve the problem

caused by interference. The work [23] considers an equidistant missing-data problem based on Granger causality. Nonetheless, the problem of missing observations caused by packet loss is still an open issue, which can result in a significant drop in inference accuracy [4]–[6], [24].

### C. Main Contributions

Addressing the need for passive topology inference techniques that are robust to packet losses, this paper introduces EM-CDA. The main contributions of this paper can be summarized as follows.

- We formulate the problem of network topology inference as the maximum likelihood problem of estimating existing network links in the presence of latent variables representing packet losses. EM-CDA is derived as a tractable approximation of the resulting EM algorithm. Accordingly, EM-CDA iterates between the estimation of packet losses and the evaluation of causality metrics based on the estimated missing packets.
- EM-CDA is validated through experiments in the wireless network on the NS-3 simulation platform, demonstrating that EM-CDA can improve the detection probability and false alarm probability rate of CDA ranging from 4% to 12% under a variety of practical conditions.

The rest of this paper is organized as follows. The wireless network scenario and system model are described in Section II. The state-of-the-art causality discovery algorithm (CDA) for wireless network topology inference is presented in Section III. EM-CDA scheme is introduced in Section IV. In Section V, numerical results are given to demonstrate the performance of the proposed algorithm. Finally, the paper is concluded in Section VI.

## II. SYSTEM MODEL AND PROBLEM SETUP

In this section, we describe the setting under study in which, as illustrated in Fig. 1, a central monitor collects meta-data about packet timings from the nodes of a network in order to infer the network topology. In this paper, unlike [4], [5], we allow packet losses to occur on the communication links. This creates additional challenges in relating the timings of data and control (acknowledgment) packets, motivating the novel estimation algorithm introduced in the next section.

### A. Setting

Consider the problem of estimating the topology of a network consisting of a set  $\mathcal{N} = \{1, 2, \dots, N\}$  of  $N$  nodes and of a set  $\mathcal{L} = \{(i, j) | i, j \in \mathcal{N}\}$  of  $M \leq N(N-1)$  *directional* links. The presence of a link  $(i, j) \in \mathcal{L}$  with  $i, j \in \mathcal{N}$  and  $i \neq j$  indicates that node  $i$  communicates with node  $j$ .

As in [4], [5], we assume that a central monitor collects meta-data in the form of transmission timestamps reporting the time instants at which data packets or acknowledgments (ACKs) are sent by each node within a given time window. Only timing meta-data is collected, and hence the monitor is only aware of packet timings, and not of the intended destination of any given packet. Successful transmission of a data packet from one node to another causes the transmission of an ACK from the receiving node to the transmitting one. ACKs are assumed to be much shorter than data packets and not subject to data losses.

### B. Data Transmission and Channel Model

The observation period  $T$  is discretized into  $K$  equal time slots of duration  $T_s = T/K$ , which are indexed by integer  $k \in \mathcal{K} = \{1, 2, \dots, K\}$ . To describe the timing information recorded by node  $i \in \mathcal{N}$ , two integer-valued time sequences  $Y_i^D[k]$  and  $Y_i^A[k]$  are introduced, corresponding to data packets and ACKs, respectively. The data packet timing sample  $Y_i^D[k]$  equals the number of data packets sent by node  $i$  in time slot  $k$ . In a similar way, the timing information sample  $Y_i^A[k]$  for ACK packets equals the number of ACK packets sent by node  $i$  in time slot  $k$ . We collect the data timing information across all time slots for node  $i$  in the  $K \times 1$  vector

$$\mathbf{Y}_i^D = [Y_i^D[1], Y_i^D[2], \dots, Y_i^D[K]]^T, \quad (1)$$

and the ACK timing information in the  $K \times 1$  vector

$$\mathbf{Y}_i^A = [Y_i^A[1], Y_i^A[2], \dots, Y_i^A[K]]^T. \quad (2)$$

The data and ACK timing series for node  $i$  can be expressed as the sum of individual contributions corresponding to the distinct communication links stemming from node  $i$ . To elaborate, we define the per-link binary sequences

$$Y_{i,j}^D[k] = \begin{cases} 1 & \text{if a data packet is sent on link } (i, j) \text{ in time slot } k, \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

and

$$Y_{i,j}^A[k] = \begin{cases} 1 & \text{if an ACK is sent on link } (i, j) \text{ in time slot } k, \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

Note that the time slot  $T_s$  is assumed to be sufficiently small so that no more than one data packet is sent by a node to another node within a single slot. Using the per-link sequences (3)-(4), the per-node observations (1)-(2) can be written as the sums

$$Y_i^D[k] = \sum_{(i,j) \in \mathcal{L}} Y_{i,j}^D[k], \quad (5)$$

and

$$Y_i^A[k] = \sum_{(i,j) \in \mathcal{L}} Y_{i,j}^A[k]. \quad (6)$$

Importantly, by collecting the sequences (5)-(6), the monitor only has aggregate information regarding the achieving of each node  $i$  while not having access to the per-link series  $Y_{i,j}^D[k]$  and  $Y_{i,j}^A[k]$ .

The data packet and ACK timing sequences are related by the ARQ protocol. Let us denote as  $\tau_{i,j}[k]$  the delay, measured in the number of time slots, between the transmission of a data packet in time slot  $k$  by node  $i$  to node  $j$  and the transmission of the corresponding ACK packets from node  $j$  to node  $i$ . We also introduce the per-link binary error variable  $E_{i,j}[k]$  defined as

$$E_{i,j}[k] = \begin{cases} 1 & \text{if an error occurs on link } (i, j) \text{ in time slot } k, \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

With these definitions, we have the equality

$$Y_{j,i}^A[k + \tau_{i,j}[k]] = (1 - E_{i,j}[k]) Y_{i,j}^D[k], \quad (8)$$

which indicates that an ACK is sent in time slot  $k + \tau_{i,j}[k]$  on link  $(j, i)$ , i.e.,  $Y_{j,i}^A[k + \tau_{i,j}[k]] = 1$ , when a data packet is sent in time slot  $k$  on the reverse link  $(i, j)$ , i.e.,  $Y_{i,j}^D[k] = 1$  and an error does not occur on link  $(i, j)$ , i.e.,  $E_{i,j}[k] = 0$ . The  $\tau_{i,j}[k]$  may severely vary across links and time slots, and it is unknown to the monitor.

### C. Topology Inference

The timing information sequences  $\{Y_i^D | \forall i \in \mathcal{N}\}$  and  $\{Y_i^A | \forall i \in \mathcal{N}\}$  in (1)-(2) collected from all nodes are used by the monitor to infer the topology, which is defined by set of links  $\mathcal{L}$ . The

links set  $\mathcal{L}$  can be equivalently also described by the adjacency matrix  $\mathbf{A} = \{a_{i,j} | \forall i, j \in \mathcal{N}\}$  with entries

$$a_{i,j} = \begin{cases} 1 & \text{if } (i, j) \in \mathcal{L}, \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

Therefore, the goal of the monitor is to use sequence  $\{\mathbf{Y}_i^D | \forall i \in \mathcal{N}\}$  and  $\{\mathbf{Y}_i^A | \forall i \in \mathcal{N}\}$  to produce an estimate  $\hat{\mathbf{A}}$  of the adjacency matrix  $\mathbf{A}$ , or equivalently an estimate  $\hat{\mathcal{L}}$  of the link set  $\mathcal{L}$ .

### III. CAUSALITY-BASED TOPOLOGY ESTIMATION

In this section, we review the Causality Discovery Algorithms (CDAs) introduced in [4]–[6], [24] wherein links are included in the estimated set  $\hat{\mathcal{L}}$  based on measures of causal dependence between data and ACK sequences of two nodes.

#### A. Causality Discovery Algorithm

In CDA schemes, the monitor estimates a measure of causal dependence  $\Phi(\mathbf{Y}_i^D \rightarrow \mathbf{Y}_j^A)$  between sequences  $\mathbf{Y}_i^D$  and  $\mathbf{Y}_j^A$  for each pair of nodes  $i$  and  $j$ . The measure  $\Phi(\mathbf{Y}_i^D \rightarrow \mathbf{Y}_j^A)$  quantifies the degree to which the future of sequence  $\mathbf{Y}_j^A$  can be predicted based on the past of sequence  $\mathbf{Y}_i^D$ . A link  $(i, j)$  is added to the estimated set  $\hat{\mathcal{L}}$  if the measure  $\Phi(\mathbf{Y}_i^D \rightarrow \mathbf{Y}_j^A)$  is larger than some threshold  $\theta_{i,j}$ . This condition can be equivalently expressed as

$$\hat{a}_{i,j} = \begin{cases} 1 & \text{if } \Phi(\mathbf{Y}_i^D \rightarrow \mathbf{Y}_j^A) > \theta_{i,j}, \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

The rationale for this decision rule is that, if link  $(i, j)$  exists, then by (8) data packets from node  $i$  cause ACKs from node  $j$ , assuming that there are no errors. This, in turn, ideally contributes to increasing the causal dependence measure  $\Phi(\mathbf{Y}_i^D \rightarrow \mathbf{Y}_j^A)$ .

We now discuss specific choice for the causal dependence measure  $\Phi(\mathbf{Y}_i^D \rightarrow \mathbf{Y}_j^A)$ .

#### B. Causality Metrics

Granger causality (GC) is a standard measure of causal dependence that is based on linear prediction. Given two time sequences  $\mathbf{Y}_i^D$  and  $\mathbf{Y}_j^A$ , GC evaluates the extent to which omitting

the past of time series  $Y_i^D[k]$  increases the prediction error for sequence  $Y_j^A[k]$  when prediction is based on a linear  $R$ -order autoregressive (AR) model. Formally, GC uses the available observations  $\mathbf{Y}_i^D$  and  $\mathbf{Y}_j^A$  to fit separately two models, namely

$$Y_j^A[k] = \sum_{r=1}^R a_{1r} Y_j^A[k-r] + \sum_{r=1}^R a_{2r} Y_i^D[k-r] + \varepsilon_k, \quad (11)$$

and

$$Y_j^A[k] = \sum_{r=1}^R b_r Y_j^A[k-r] + \eta_k, \quad (12)$$

by optimising over parameters  $\{a_{1r}, a_{2r}\}_{r=1}^R$ , and  $\{b_r\}_{r=1}^R$  via least squares minimization. In (11)-(12), the quantities  $\varepsilon_k$  and  $\eta_k$  represent the prediction residuals. The prediction residuals  $\varepsilon_k$  in (11) account for prediction errors accrued on the ACK sequence  $Y_j^A[k]$  when the past of data packet sequence  $Y_i^D[k]$  is known; while the residuals  $\eta_k$  in (12) are obtained when prediction can only use the past sample for the ACK sequence  $Y_j^A[k]$  itself. The GC-based measure is given by [6]

$$\Phi_{\text{GC}}(\mathbf{Y}_i^D \rightarrow \mathbf{Y}_j^A) = \frac{(\sum_{k=1}^H |\eta_k|^2 - \sum_{k=1}^H |\varepsilon_k|^2)/R}{\sum_{k=1}^H |\varepsilon_k|^2/(K-3R-1)}, \quad (13)$$

which is large when the sum-residual  $\sum_{k=1}^H |\eta_k|^2$  is larger than  $\sum_{k=1}^H |\varepsilon_k|^2$ , where  $H = K - R$ . GC was used in [6], [24] for topology estimation.

Transfer entropy (TE) is an information-theoretic causality measure that does not assume a linear relation between sequences  $\mathbf{Y}_i^D$  and  $\mathbf{Y}_j^A$  as GC. To introduce it, let us define as  $I(A; B|C)$  the conditional mutual information of random variable  $A$  and  $B$  given  $C$ , which is defined as

$$I(A; B|C) = \mathbb{E} \left[ \log_2 \frac{p(A|B, C)}{p(A|C)} \right], \quad (14)$$

where the expectation is taken over the point distribution  $p(A|B, C)$  and  $p(A|C)$ . With these definitions, the TE is defined as [12]

$$\begin{aligned} \Phi_{\text{TE}}(\mathbf{Y}_i^D \rightarrow \mathbf{Y}_j^A) \\ = I(Y_j^A[k]; \mathbf{Y}_i^D[k-1:k-s] | \mathbf{Y}_j^A[k-1:k-r]), \end{aligned} \quad (15)$$

where  $s$  and  $r$  are fixed integers;  $\mathbf{Y}_j^A[k-1:k-r] = \{Y_j^A[k-1], Y_j^A[k-2], \dots, Y_j^A[k-r]\}$  and  $\mathbf{Y}_i^D[k-1:k-s] = \{Y_i^D[k-1], Y_i^D[k-2], \dots, Y_i^D[k-s]\}$  denote windows of past samples for  $\mathbf{Y}_j^A$  and  $\mathbf{Y}_i^D$  respectively. In practice, the TE is estimated using available data sequences  $\mathbf{Y}_j^A$  and  $\mathbf{Y}_i^D$ . The TE was used for topology estimation in [4], [5]

### C. Setting the Threshold

The threshold  $\theta_{i,j}$  in (10) can be set via a permutation test [25]. Accordingly, one considers a statistical significance test in which the null hypothesis corresponds to the assumption that the two sequences  $\mathbf{Y}_i^D$  and  $\mathbf{Y}_j^A$  are not causally related. To obtain the distribution of the causality metrics  $\Phi(\mathbf{Y}_i^D \rightarrow \mathbf{Y}_j^A)$  under the null hypothesis,  $S$  random permutations of the sequences are obtained by considering permutations of the observed sequences.  $\mathbf{Y}_{i,s}^D$  and  $\mathbf{Y}_{j,s}^A$  of sequences  $\mathbf{Y}_i^D$  and  $\mathbf{Y}_j^A$  are produced, with  $s \in \{1, 2, \dots, S\}$ . The causality metrics  $\Phi(\mathbf{Y}_{i,s}^D \rightarrow \mathbf{Y}_{j,s}^A)$ , with  $s \in \{1, 2, \dots, S\}$ , are evaluated; and the threshold  $\theta_{i,j}$  is set as the  $(1 - \alpha)$ -quantile of the empirical distribution of the samples  $\{\Phi(\mathbf{Y}_{i,s}^D \rightarrow \mathbf{Y}_{j,s}^A)\}_{s=1}^S$ , when  $\alpha \in [0, 1]$  is a fixed false alarm probability.

## IV. EM-BASED TOPOLOGY ESTIMATION

The CDA schemes reviewed in the previous section were devised under the assumption that there are no packet losses [4]–[6]. As we argue in Sec. IV-A, packet losses tend to make the CDA test (10) unreliable, since the causality metrics are decreased in the presence of packet losses due to the missed association between data and ACK sequences erased by lost data packets. To address this challenge, in this section, we introduce the EM-based CDA, which models packet losses using latent random variables.

### A. Impact of Packet Losses on CDA

In order to gain insights into the impact of packet losses on the performance of CDA, we now consider an IEEE 802.11 ad-hoc network simulated with NS-3, and evaluate the GC metric (13) for a given link  $(i, j)$  in the presence and absence of packet losses. Details of the experimental setting can be found in Sec. V-B. Fig. 2 reports the GC metric evaluated with losses as a function of the corresponding metric evaluated in a lossless scenario under the same conditions. Different points correspond to distinct links in the set  $\mathcal{L}$ . The figure confirms that the GC metric tends to be decreased by packet losses, making CDA methods potentially ineffective.

### B. Parametric Model with Latent Variables

The EM-Based Causality Discovery Algorithm (EM-CDA) scheme is based on the idea of formulating the problem of topology inference as the maximum likelihood estimate (MLE) of the adjacency matrix  $\mathbf{A}$  in the presence of latent variables describing packet losses. To elaborate,

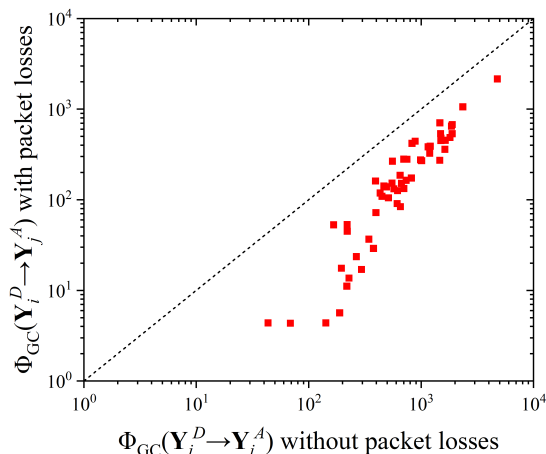


Fig. 2. Causality metric (13) evaluated for different links of an IEEE 802.11 ad-hoc network simulated on NS-3, with  $N = 12$  nodes,  $M = 65$  links, observation duration 60 s, time slot duration  $T_s = 1.5$  ms, and probability of packet loss 0.25.

let  $\mathbf{Y} = \{\mathbf{Y}_i^D, \mathbf{Y}_i^A | \forall i \in \mathcal{N}\}$  be the observations. We also introduce two sets of latent variables. The first,  $\mathbf{D} = \{D_{i,j}[k] | \forall i, j \in \mathcal{N}, k \in \mathcal{K}\}$ , contains variables  $D_{i,j}[k]$  for all pairs of nodes  $i$  and  $j$  and time slots  $k$ , such that

$$D_{i,j}[k] = \begin{cases} 1 & \text{if a data packet is sent on link } (i, j) \text{ in time slot } k, \\ 0 & \text{otherwise.} \end{cases} \quad (16)$$

The second,  $\mathbf{E} = \{E_{i,j}[k] | \forall i, j \in \mathcal{N}, k \in \mathcal{K}\}$ , contains the packet loss variables defined in (7). Note that the true value of the latent variables  $D_{i,j}[k]$  and  $E_{i,j}[k]$  are undefined for links not in set  $\mathcal{L}$ . The set  $\mathbf{Z} = \{\mathbf{E}, \mathbf{D}\}$  defines the latent variables. Overall, we have observations  $\mathbf{Y}$  and latent variables  $\mathbf{Z}$ .

We now define a parametric model that specifies the point distribution  $p(\mathbf{Y}, \mathbf{Z} | \Theta)$  of observations  $\mathbf{Y}$  and latent variables  $\mathbf{Z}$  as a function of a set of parameters,  $\Theta$ . Set  $\Theta$  includes the adjacency matrix  $\mathbf{A}$ , which is the quantity of interest, as well as some nuisance parameters to be introduced next. We emphasize that the probabilistic model  $p(\mathbf{Y}, \mathbf{Z} | \Theta)$  does not generally describe the ground-truth data generation mechanism, which is unknown. Rather, it amounts to a set of assumptions made in order to develop the proposed topology estimation algorithm.

The parametric model,  $p(\mathbf{Y}, \mathbf{Z} | \Theta) = p(\mathbf{Z} | \Theta)p(\mathbf{Y} | \mathbf{Z}; \Theta)$ , depends on the set of unknown parameters  $\Theta = \{\mathbf{A}, \mathbf{L}, \mathbf{R}, \mathbf{T}\}$ , where matrices  $\{\mathbf{L}, \mathbf{R}, \mathbf{T}\}$  are nuisance parameter matrices representing error rate, transmission rate, and ACK delay on each link, respectively. Let us

define as  $\mathcal{E}(\mathbf{A})$  the set of coordinates of non-zero entries of the adjacency matrix  $\mathbf{A}$ , that is, the estimated links given matrix  $\mathbf{A}$ . To start, we assume that variables  $(E_{i,j}[k], D_{i,j}[k])$  corresponding to different link  $(i, j)$  are independent, i.e.,

$$p(\mathbf{Z}|\Theta) = \prod_{(i,j) \in \mathcal{E}(\mathbf{A})} p(\mathbf{E}_{i,j}, \mathbf{D}_{i,j}|\Theta), \quad (17)$$

where we have the sequences  $\mathbf{E}_{i,j} = \{E_{i,j}[k]\}_{k=1}^K$  and  $\mathbf{D}_{i,j} = \{D_{i,j}[k]\}_{k=1}^K$ . Focusing now on sequences  $\mathbf{E}_{i,j}$  and  $\mathbf{D}_{i,j}$ , we assume the joint distribution

$$\begin{aligned} p(\mathbf{E}_{i,j}, \mathbf{D}_{i,j}|\Theta) &= \prod_{k \in \mathcal{K}} \left[ p(D_{i,j}[k]|\mathbf{D}_{i,j}[1:k-1], \mathbf{E}_{i,j}[1:k-1]; \Theta) \right. \\ &\quad \left. \times p(E_{i,j}[k]|\mathbf{D}_{i,j}[1:k], \mathbf{E}_{i,j}[1:k-1]; \Theta) \right] \\ &= \prod_{k \in \mathcal{K}} \left[ p(D_{i,j}[k]|R_{i,j}) p(E_{i,j}[k]|D_{i,j}[k]; L_{i,j}) \right], \end{aligned} \quad (18)$$

where the first equality follows from the chain rule of probability, and the second is a consequence of the following two assumptions. First, we assume that an error on a link  $(i, j)$ , indicated by  $E_{i,j}[k] = 1$ , occurs with probability  $L_{i,j}$  if a transmission occurred on the same link, i.e., if  $D_{i,j}[k] = 1$ . This is expressed with the conditional distribution

$$\begin{aligned} &p(E_{i,j}[k]|D_{i,j}[k]; L_{i,j}) \\ &= \begin{cases} L_{i,j}^{E_{i,j}[k]} (1 - L_{i,j})^{1-E_{i,j}[k]} & \text{if } D_{i,j}[k] = 1, \\ 1 - E_{i,j}[k] & \text{otherwise.} \end{cases} \end{aligned} \quad (19)$$

Second, transmissions occur independently of previous transmissions and errors with probability  $R_{i,j}$ , which is formulated as

$$p(D_{i,j}[k]|R_{i,j}) = R_{i,j}^{D_{i,j}[k]} (1 - R_{i,j})^{1-D_{i,j}[k]}. \quad (20)$$

We emphasize that the conditional distribution (20) entails a significant approximation, since transmissions in many network scenarios encompass also retransmission of previous, erroneously received, packets. The Bayesian network that describes the assumed model for the latent variables is shown in Fig. 3.

To fully specify the parametric model  $p(\mathbf{Y}, \mathbf{Z}|\Theta) = p(\mathbf{Z}|\Theta)p(\mathbf{Y}|\mathbf{Z}; \Theta)$  we need to describe also the distribution  $p(\mathbf{Y}|\mathbf{Z}; \Theta)$ . In this regard, the observations  $\mathbf{Y}$  are assumed to be a function  $f(\mathbf{D}, \mathbf{E}|\mathbf{A}, \mathbf{T})$  of the latent variables  $\mathbf{D}$  and  $\mathbf{E}$  that is parameterized by the adjacency matrix  $\mathbf{A}$  and the matrix of delays  $\mathbf{T}$ . Accordingly, the distribution of  $\mathbf{Y}$  conditioned on  $\mathbf{Z}$  is given by

$$p(\mathbf{Y}|\mathbf{Z}; \Theta) = p(\mathbf{Y}|\mathbf{D}, \mathbf{E}; \mathbf{A}, \mathbf{T}) = \delta(\mathbf{Y} - f(\mathbf{D}, \mathbf{E}|\mathbf{A}, \mathbf{T})), \quad (21)$$

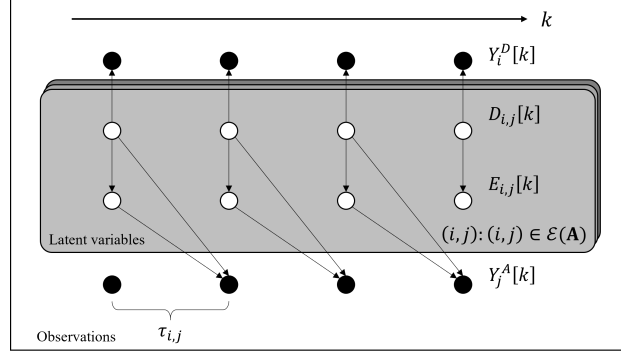


Fig. 3. Bayesian network of the parametric model assumed in the derivation of EM-CDA. Shaded circles correspond to observed variables, and we set  $\tau_{i,j} = 1$  for simplicity of illustration. Note that the observations  $Y_i^D[k]$  and  $Y_j^A[k]$  depend only on latent variables indexed by  $i$  and  $j$ , respectively, with  $(i, j) \in \mathcal{E}(\mathbf{A})$ .

where  $\delta(\cdot)$  is the Kronecker delta function. Function  $f(\mathbf{D}, \mathbf{E} | \mathbf{A}, \mathbf{T})$  is defined as follows. Since the number of packets observed from a node  $i$  equals the sum of the numbers of packets sent to other nodes  $j$  with  $(i, j) \in \mathcal{E}(\mathbf{A})$  in the given time slot, we have the equality

$$Y_i^D[k] = \sum_{j:(i,j) \in \mathcal{E}(\mathbf{A})} D_{i,j}[k]. \quad (22)$$

This is reflected by the Bayesian network in Fig. 3. Similarly, the number of ACKs reported by a node  $j$  is equal to the sum of the numbers of ACKs sent to other nodes  $i$  with  $(i, j) \in \mathcal{E}(\mathbf{A})$ . Defining the model parameter  $\tau_{i,j}$  as the delay between ACK and packet transmission on link  $(i, j)$ , we thus assume the equality

$$Y_j^A[k] = \sum_{i:(i,j) \in \mathcal{E}(\mathbf{A})} (1 - E_{i,j}[k - \tau_{i,j}]) D_{i,j}[k - \tau_{i,j}]. \quad (23)$$

It is recalled that, while the actual unknown time delays  $\tau_{i,j}[k]$  in (5) may depend on the time slot  $k$ , the parameters  $\tau_{i,j}$  in (23), which are collected in matrix  $\mathbf{T}$ , are assumed to be static in order to facilitate estimation. Overall, equalities (22)-(23) define function  $f(\mathbf{D}, \mathbf{E} | \mathbf{A}, \mathbf{T})$  and hence distribution (21).

### C. EM-Based Causality Discovery Algorithm (EM-CDA)

Given the likelihood  $p(\mathbf{Y}, \mathbf{Z} | \Theta)$  of the complete data  $(\mathbf{Y}, \mathbf{Z})$ , EM-CDA aims to address the MLE problem

$$\max_{\Theta} \{p(\mathbf{Y} | \Theta) = \mathbb{E}_{p(\mathbf{Z} | \Theta)} [p(\mathbf{Y} | \mathbf{Z}; \Theta)]\} \quad (24)$$

via EM. Accordingly, EM-CDA updates the current estimate  $\Theta$  across a number of iteration, producing a sequence of iterates  $\Theta^{(1)}, \Theta^{(2)}, \dots, \Theta^{(n)}$ . At each iteration  $n$ , EM first performs the expectation step (E-step), which evaluates the expected value

$$Q(\Theta|\Theta^{(n)}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{Y};\Theta^{(n)})} [\log p(\mathbf{Y}, \mathbf{Z}|\Theta)] \quad (25)$$

of the complete log-likelihood  $\log p(\mathbf{Y}, \mathbf{Z}|\Theta)$  with respect to the current posterior distribution  $p(\mathbf{Z}|\mathbf{Y}; \Theta^{(n)})$ . Then, the maximization step (M-step) is carried out, wherein the next update is obtained as

$$\Theta^{(n+1)} = \underset{\Theta}{\operatorname{argmax}} Q(\Theta|\Theta^{(n)}). \quad (26)$$

A direct application of EM to the model  $p(\mathbf{Y}, \mathbf{Z}|\Theta)$  described in the previous subsection is computationally infeasible. To obtain a scalable solution, EM-CDA approximates the E-step using Monte Carlo sampling, and the M-step via CDA (see Sec. III). The resulting algorithm can be viewed as an iterative generalization of CDA, wherein estimates of packet losses are accounted for in the estimates of the causality metrics in order to address the issue described in Sec. IV-A. We detail both E-step and M-step in the rest of this section, and the overall EM-CDA is described in Algorithm 1.

#### D. Expectation Step (E-step)

At iteration  $n$ , given the current parameters  $\Theta^{(n)}$ , the E-step aims at generating  $M$  samples  $\{\mathbf{Z}_1^{(n)}, \dots, \mathbf{Z}_M^{(n)}\}$  from the posterior distribution  $p(\mathbf{Z}|\mathbf{Y}; \Theta^{(n)})$ . With such samples, the function  $Q(\Theta|\Theta^{(n)})$  in (25) is approximated via the stochastic estimate [26]

$$\begin{aligned} Q(\Theta|\Theta^{(n)}) &= (1 - \gamma^{(n)}) Q(\Theta|\Theta^{(n-1)}) \\ &\quad + \frac{\gamma^{(n)}}{M} \sum_{m=1}^M \log p(\mathbf{Y}, \mathbf{Z}_m^{(n)}|\Theta), \end{aligned} \quad (27)$$

where  $\gamma^{(n)} \in [0, 1]$  is a learning rate.

In order to generate the samples  $\mathbf{Z}_m^{(n)} \sim p(\mathbf{Z}|\mathbf{Y}; \Theta^{(n)})$  for  $m = 1, \dots, M$ , we apply Gibbs sampling. Gibbs sampling generates the samples  $\mathbf{Z}_m^{(n)}$  sequentially over index  $m = 1, \dots, M$  by drawing samples from the conditional probabilities of one variable in  $\mathbf{Z}$  given all other variables in  $\mathbf{Z}$  [27]. Accordingly, each sample  $\mathbf{Z}_m^{(n)} = \{D_{i,j,m}^{(n)}[k], E_{i,j,m}^{(n)}[k] | \forall i, j \in \mathcal{N}, k \in \mathcal{K}\}$  is generated as follows.

Using the notations  $Z_{i,j}[k] = (D_{i,j}[k], E_{i,j}[k])$  and  $Z_{-(i,j)}[-k] = \{D_{i',j'}[k], E_{i',j'}[k]\}_{(i',j') \neq (i,j), k' \neq k}$ , for each pair of variables  $Z_{i,j}[k]$ , we sample from the posterior  $p(Z_{i,j}[k] | Z_{-(i,j)}[-k], \mathbf{Y}; \Theta^{(n)})$  given all other variables. This can be evaluated as

$$\begin{aligned}
& p(Z_{i,j}[k] | Z_{-(i,j)}[-k], \mathbf{Y}; \Theta^{(n)}) \\
&= p(Z_{i,j}[k] | Y_i^D[k], Y_j^A[k + \tau_{i,j}]; \Theta^{(n)}) \\
&= \frac{p(Z_{i,j}[k], Y_i^D[k], Y_j^A[k + \tau_{i,j}] | \Theta^{(n)})}{p(Y_i^D[k], Y_j^A[k + \tau_{i,j}] | \Theta^{(n)})} \\
&= \frac{p(Z_{i,j}[k] | \Theta^{(n)}) p(Y_i^D[k], Y_j^A[k + \tau_{i,j}] | Z_{i,j}[k]; \Theta^{(n)})}{\sum_{Z_{i,j}[k]} p(Z_{i,j}[k] | \Theta^{(n)}) p(Y_i^D[k], Y_j^A[k + \tau_{i,j}] | Z_{i,j}[k]; \Theta^{(n)})}, \tag{28}
\end{aligned}$$

where the first equality follows from d-separation based on the Bayesian network in Fig. 3 (see, e.g., [28]), and  $p(Z_{i,j}[k] | \Theta^{(n)})$  is given by the product of (19) and (20) as

$$p(Z_{i,j}[k] | \Theta^{(n)}) = p(D_{i,j}[k] | R_{i,j}^{(n)}) p(E_{i,j}[k] | D_{i,j}[k]; L_{i,j}^{(n)}). \tag{29}$$

We now left with the problem evaluating the distribution  $p(Y_i^D[k], Y_j^A[k + \tau_{i,j}] | Z_{i,j}[k]; \Theta^{(n)})$ . According to (22)-(23), it is given by the probability of that  $Y_i^D[k] - D_{i,j}[k]$  packets are sent by node  $i$  to other nodes except  $j$  at time  $k$ , and that  $Y_j^A[k + \tau_{i,j}] - D_{i,j}[k](1 - E_{i,j}[k])$  ACKs are sent by node  $j$  to other nodes except  $i$  at time  $k + \tau_{i,j}$ . Therefore, by (18)-(20) we have

$$\begin{aligned}
& p(Y_i^D[k], Y_j^A[k + \tau_{i,j}] | Z_{i,j}[k]; \Theta^{(n)}) \\
&= p(Y_i^D[k] | Z_{i,j}[k]; \Theta^{(n)}) p(Y_j^A[k + \tau_{i,j}] | Z_{i,j}[k]; \Theta^{(n)}) \\
&= \text{Bin}\left(Y_i^D[k] - D_{i,j}[k] | \{R_{i,j}^{(n)}\}_{\substack{(i,l) \in \mathcal{E}(\mathbf{A}) \\ l \neq j}}}\right) \\
&\quad \times \text{Bin}\left(Y_j^A[k + \tau_{i,j}] - D_{i,j}[k](1 - E_{i,j}[k]) | \{R_{i,j}^{(n)}(1 - L_{i,j}^{(n)})\}_{\substack{(l,j) \in \mathcal{E}(\mathbf{A}) \\ l \neq i}}}\right), \tag{30}
\end{aligned}$$

where we denote as  $\text{Bin}(y | \{p_i\}_{i=1}^L)$  the probability mass function of a sum of  $L$  independent Bernoulli random variables, with each  $i$ th random variables having probability  $p_i$  of being equal to 1.

### E. Maximization Step (M-step)

Given  $\mathbf{Y}$  and samples generated  $\{\mathbf{Z}_1^{(n)}, \dots, \mathbf{Z}_M^{(n)}\}$  in the E-step, the M-step aims at updating parameters  $\Theta$ . The discrete parameters  $\mathbf{T}$  and  $\mathbf{A}$  are updated by generalizing the CDA approach described in Sec. III to include the estimate of delays. The continuous parameters  $\mathbf{R}$  and  $\mathbf{L}$  are then updated by finding the stationary points of the objective function of  $Q(\Theta | \Theta^{(n)})$  in (27).

For each sample  $\mathbf{Z}_m^{(n)}$ , we define as  $\mathbf{Y}_{i,m}^{D,(n)} = \left\{ \sum_{j \in \mathcal{N}} D_{i,j,m}^{(n)}[k] \mid \forall k \in \mathcal{K} \right\}$  the estimated data packet sequence for node  $i$ ; and as  $\mathbf{Y}_{j,m}^{A,(n)} = \left\{ Y_j^A[k] + \sum_{(i,j) \in \mathcal{E}(\mathbf{A})} E_{i,j,m}^{(n)}[k - \tau_{i,j}^{(n)}] \mid \forall k \in \mathcal{K} \right\}$  the estimated ACK sequence for node  $j$ . To update the delay matrix for  $\mathbf{T}_m^{(n)} = \left\{ \tau_{i,j,m}^{(n)} \mid \forall i, j \in \mathcal{N} \right\}$ , we obtain the sequences  $\mathbf{Y}_{i,m}^{D,(n),\tau} = \{Y_{i,m}^{D,(n)}[k + \tau]\}_{k=1}^K$  by shifting backward in time by  $\tau$  steps the sequences  $\mathbf{Y}_{i,m}^{D,(n)}$ . Then, the causal dependence measure  $\Phi(\mathbf{Y}_{i,m}^{D,(n),\tau} \rightarrow \mathbf{Y}_{j,m}^{A,(n)})$  is calculated using (13) or (15) for a range of values  $[1, \tau_{max}]$  to obtain the estimate

$$\tau_{i,j,m}^{(n)} = \operatorname{argmax}_{\tau \in [1, \tau_{max}]} \Phi \left( \mathbf{Y}_{i,m}^{D,(n),\tau} \rightarrow \mathbf{Y}_{j,m}^{A,(n)} \right). \quad (31)$$

Furthermore, using (10), the estimated topology entries  $a_{i,j,m}^{(n)}$  of the adjacency matrix  $\mathbf{A}_m^{(n)}$  are given by

$$a_{i,j,m}^{(n)} = \begin{cases} 1 & \text{if } \Phi(\mathbf{Y}_{i,m}^{D,(n)} \rightarrow \mathbf{Y}_{j,m}^{A,(n)}) > \theta_{i,j,m}^{(n)}, \\ 0 & \text{otherwise,} \end{cases} \quad (32)$$

where  $\theta_{i,j,m}^{(n)}$  is a threshold. Then, we set

$$a_{i,j}^{(n+1)} = \begin{cases} 1 & \text{if } \sum_{m=1}^M a_{i,j,m}^{(n)} \geq \frac{M}{2}, \\ 0 & \text{otherwise,} \end{cases} \quad (33)$$

that is, an edge  $(i, j)$  is included in the set  $\mathcal{E}(\mathbf{A}^{(n)})$  of the majority of tests (32) set  $a_{i,j,m}^{(n)} = 1$ .

Finally, setting the partial derivatives of  $Q(\Theta | \Theta^{(n)})$  in (27) with respect to  $\mathbf{R}$  and  $\mathbf{L}$  to zero, respectively, the updated  $\mathbf{R}^{(n+1)}$  and  $\mathbf{L}^{(n+1)}$  are given by the empirical averages as

$$R_{i,j}^{(n+1)} = (1 - \gamma^{(n)}) R_{i,j}^{(n)} + \frac{\gamma^{(n)}}{MK} \sum_{m=1}^M \sum_{k \in \mathcal{K}} D_{i,j,m}^{(n)}[k], \quad \forall (i, j) \in \mathcal{E}(\mathbf{A}^{(n)}), \quad (34)$$

and

$$L_{i,j}^{(n+1)} = (1 - \gamma^{(n)}) L_{i,j}^{(n)} + \gamma^{(n)} \frac{\sum_{m=1}^M \sum_{k \in \mathcal{K}} E_{i,j,m}^{(n)}[k]}{\sum_{m=1}^M \sum_{k \in \mathcal{K}} D_{i,j,m}^{(n)}[k]}, \quad \forall (i, j) \in \mathcal{E}(\mathbf{A}^{(n)}). \quad (35)$$

## V. NUMERICAL RESULTS

In this section, numerical results are provided to demonstrate the performance of the proposed EM-CDA scheme as corresponds to the conventional CDA methods reviewed in Sec. III [4]–[6]. We first consider a toy example in which we can evaluate the impact of the approximations adopted in the derivation of EM-CDA via an exact implementation of EM. Then, large-scale experiments are conducted by simulating wireless networks via NS-3 [29], [30].

---

**Algorithm 1** EM-CDA
 

---

**Input:** The observations  $\mathbf{Y}$ , learning rate sequences  $\{\gamma^{(n)}\}$ , number of samples  $M$ , maximum estimated delay  $\tau_{max}$ , and significance level  $\alpha$ ;

**Output:** Matrix of inferred communication links  $\hat{\mathbf{A}}$ ;

- 1: **Initialization:** Initialize  $\mathbf{L}^{(0)}$  and  $\mathbf{R}^{(0)}$  with 0-1 uniform distribution; the adjacency matrix  $\mathbf{A}^{(0)}$  to have every entry equal to one;  $\mathbf{T}^{(0)}$  by (31) using  $\mathbf{Y}$ ;  $n$  to 0;
  - 2: **while**  $\Theta^{(n)}$  has not converged **do**
  - 3:   Generate samples based on (28)-(30);
  - 4:   Update  $\Theta^{(n+1)}$  based on (31)-(35);
  - 5:    $n \leftarrow n + 1$ ;
  - 6: **end while**
  - 7: Obtain  $\hat{\mathbf{A}} = \mathbf{A}^{(n)}$ .
- 

### A. Small-Scale Experiments

In this subsection, we compare EM-CDA with an implementation of EM to address the MLE problem (24) that applies the exhaustive search (ES) method in the M-step to maximize the function  $Q(\Theta|\Theta^{(n)})$  over variables  $\mathbf{A}$  and  $\mathbf{T}$ . We refer to this scheme as EM-ES. To enable EM-ES over the exponential number of possible choices  $\mathbf{A}$ , we consider a small network with  $N = 4$  nodes that is allowed to follow the same model adopted for the derivation of EM as explained in Sec. IV. In the next subsection, we will consider a more realistic scenario in NS-3.

Half of the links are randomly selected to be active; the ground-truth average transmission rate  $R_{i,j}^*$  for all active links is set as 0.1; the average packet loss rate  $L_{i,j}^*$  for all links is set to 0.05 or 0.5; and the ground-truth delay  $\tau_{i,j}^*$  are set to 1 time slot. We simulate the network for 5000 time slots. In the E-steps of both methods, the number of samples is set as  $M = 30$ . In M-step, GCT or TE is adopted as the causality discovery algorithm in EM-CDA. The significance level  $\alpha$  in (10) is set to 0.05 as in [5]. All the results are generated in 20 trials with different random initial values.

The probability of false alarm,  $p_{FA}$ , and the probability of detection,  $p_D$ , are adopted to measure the performance of topology inference. These metrics are defined as

$$P_{FA} = \frac{FP}{FP + TN}, \quad (36)$$

and

$$P_D = \frac{TP}{TP + FN}, \quad (37)$$

where TP denotes the number of correctly detected existing links, FN denotes the number of missed existing links, TN denotes the number of correctly detected missing links, and FP denotes the number of incorrectly detected missing links.

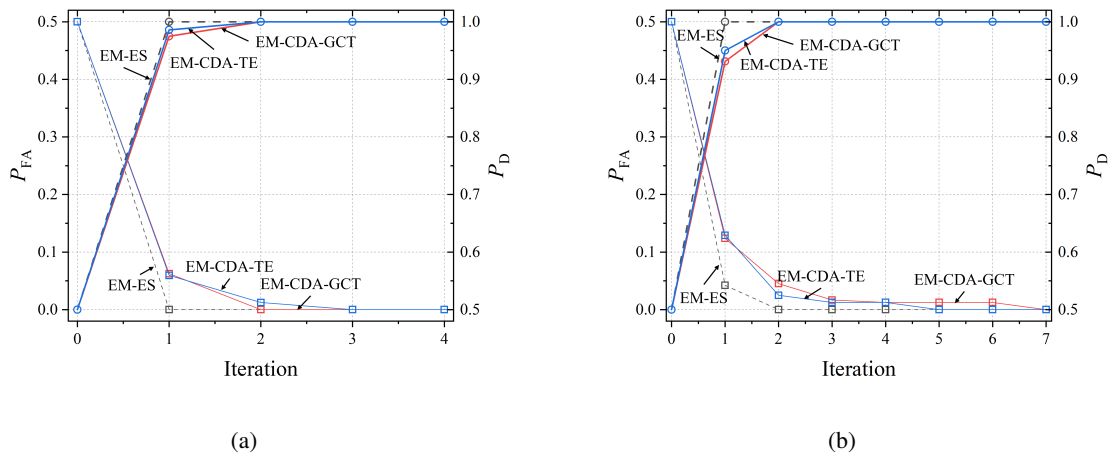


Fig. 4. Probability of false alarm  $P_{FA}$  and probability of detection  $p_D$  for topology inference versus the number of EM iterations for the ideal EM-ES scheme and for EM-CDA with GCT and TE causality metrics: iterations in the case of (a)  $L_{i,j}^* = 0.05$ , and (b)  $L_{i,j}^* = 0.5$ .

Fig. 4 shows the probabilities  $P_{FA}$  and  $P_D$  across the EM iterations. EM-ES is seen to obtain the optimal solution, yielding the ideal case  $P_{FA} = 0$  and  $P_D = 1$ , in a single iteration, while EM-CDA with both GCT and TE requires more iterations, but it is able to converge to the optimal solution. Furthermore, the number of required EM iterations for the performance of EM-CDA increases as the ground-truth average loss rate  $L_{i,j}^*$  increases, because, as discussed in Sec. IV-A, unreliable observations provide missing and spurious information that needs to be compensated for by refining the estimates of the latent variables.

### B. Simulations on NS-3

In this subsection, we test EM-CDA in different wireless scenarios simulated on NS-3 using the parameters in Table I. The system consists of  $N$  nodes randomly and uniformly distributed within a  $10 \text{ m} \times 10 \text{ m}$  area that follow the an IEEE 802.11 ad-hoc protocol operating at carrier

TABLE I  
PARAMETERS VALUES FOR NS-3 SIMULATIONS

Parameter	Value
Area size	100 m <sup>2</sup>
Carrier frequency $f_0$	2.412 GHz
Data packet size	1024 Bytes
MAC ACK size	36 Bytes
Channel packet loss rate	varies
Transmission rate	varies
Simulation duration $T$	60 s
Time slot duration $T_s$	1.5 ms

frequency  $f_0 = 2.412$  GHz. Omnidirectional antennas are used at the nodes, with path-loss, log-normal shadowing, and thermal noise accounted for as in [29]. The simulation lasts  $T = 60$  s, and the time slot duration is  $T_s = 1.5$  ms. The offered traffic for each link is 1 Mbps, with a data packet size of 1024 Bytes and an ACK size of 36 Bytes. If not stated otherwise, we set  $N = 12$  nodes, fraction of active links 0.5, and average packet loss rate 0.3.

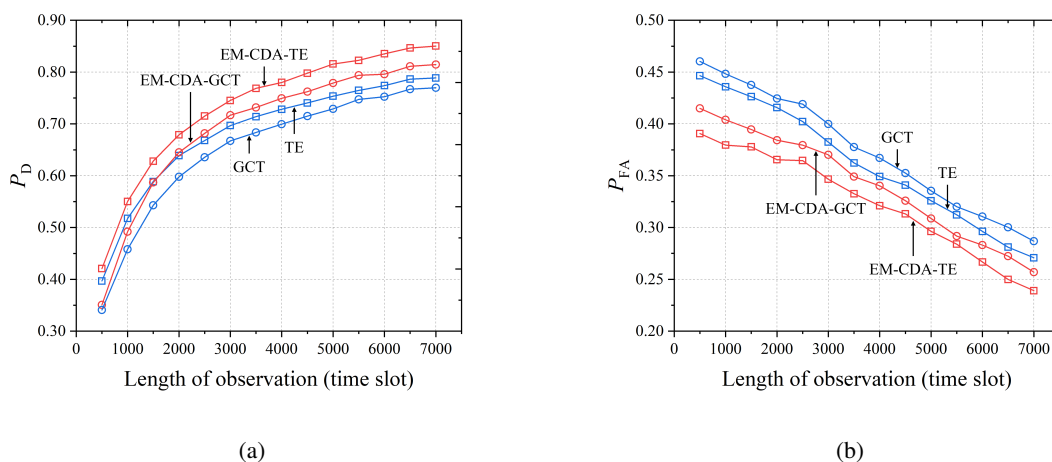


Fig. 5. (a) Probability of detection  $p_D$  and (b) probability of false alarm  $P_{FA}$  for topology inference versus the length of observation in time slots for CDA and EM-CDA with GCT and TE causality metrics.

Fig. 5 depicts the probabilities  $P_{FA}$  and  $P_D$  as a function of the number of observed time slots. As more data are collected, EM-CDA is able to outperform CDA methods in terms of

both probabilities, with gains saturating when enough information is collected.

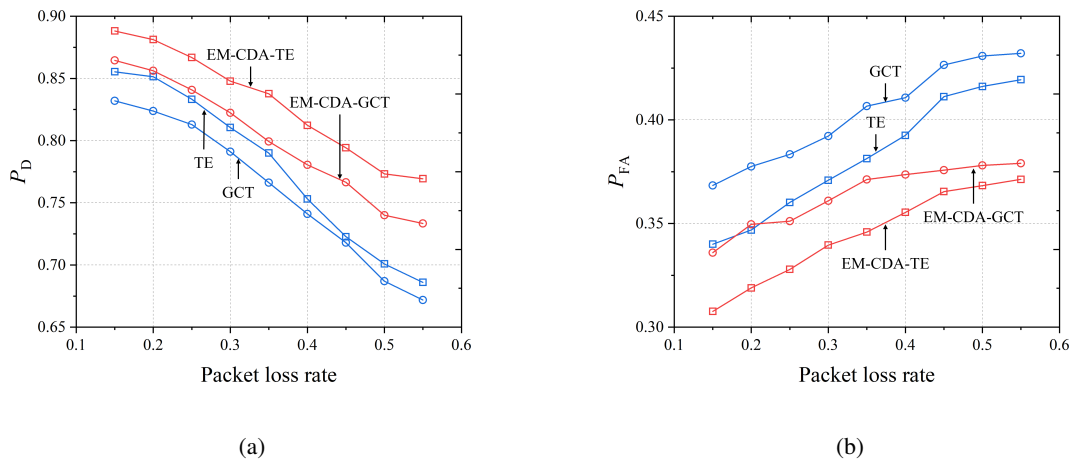


Fig. 6. (a) Probability of detection  $p_D$  and (b) probability of false alarm  $P_{FA}$  for topology inference versus the packet loss rate for CDA and EM-CDA with GCT and TE causality metrics.

The performance of CDA and EM-CDA is investigated as a function of the ground-truth packet loss rate in Fig. 6. It is shown that the detection probability of the CDA schemes decreases as the packet loss rate increases, while the false alarm probability increases. EM-CDA is seen to be able to compensate for some of this performance loss, especially when using TE.

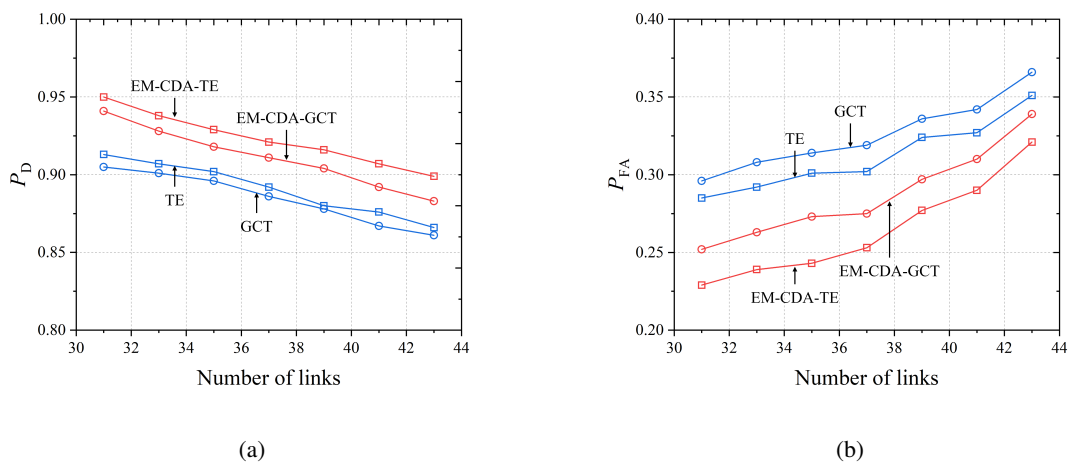


Fig. 7. (a) Probability of detection  $p_D$  and (b) probability of false alarm  $P_{FA}$  for topology inference versus the number of links for CDA and EM-CDA with GCT and TE causality metrics.

The relation between inference performance and the active link is investigated in Fig. 7, while the number of nodes is fixed as  $N = 10$ , and the number of the active link is changed from 31

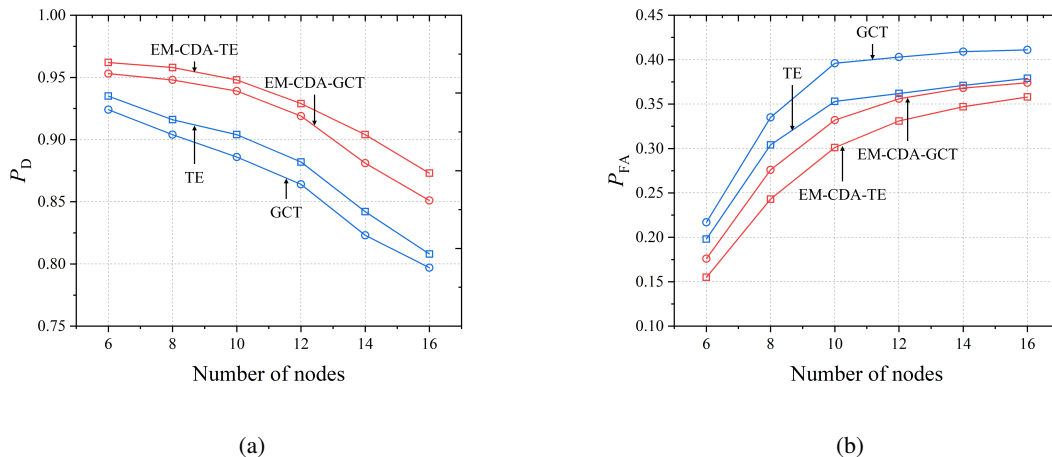


Fig. 8. (a) Probability of detection  $p_D$  and (b) probability of false alarm  $P_{FA}$  for topology inference versus the number of nodes for CDA and EM-CDA with GCT and TE causality metrics.

to 43. With an increase in the active link, the mutual interference between nodes gets larger, but EM-CDA is able to retain its performance advantage as compared to CDA method. A similar conclusion is reached from Fig. 8, which varies the number of nodes  $N$  for a fixed fraction, 0.3, of active links.

## VI. CONCLUSION

In this paper, we have introduced EM-CDA, a novel algorithm for passive network topology inference based on the observation of timing meta-data. The approach builds on the state-of-the-art causality discovery algorithm (CDA), and it addresses the important open problem of mitigating the effect of packet losses. Packet losses cause some of the timings of data packets to have no ACK packet counterparts, making CDA schemes potentially ineffective. EM-CDA formulates the topology inference problem as the discrete maximum likelihood (ML) problem of identifying active links in the presence of latent packet losses. It alternates between estimation of packet losses and application of a CDA strategy. Numerical results based on NS-3 simulations of real-world networks show that EM-CDA outperforms CDA in terms of detection probability and false alarm probability by a range of 4% to 12% under a variety of network conditions accounting for different packet loss rates, number of nodes, and active links. Future work may investigate more accurate approximations of the EM algorithm, e.g., in the evaluation of the

posterior distribution in the E step, as well as the adoption of a more detailed model to define the ML problem.

## REFERENCES

- [1] M. Cociglio, G. Fioccola, G. Marchetto, A. Sapiro, and R. Sisto, "Multipoint passive monitoring in packet networks," *IEEE/ACM Trans. Netw.*, vol. 27, no. 6, pp. 2377–2390, 2019.
- [2] P.-O. Brissaud, J. Franc¸ois, I. Chrisment, T. Cholez, and O. Bettan, "Transparent and service-agnostic monitoring of encrypted web traffic," *IEEE Trans. Netw. Serv. Manage.*, vol. 16, no. 3, pp. 842–856, 2019.
- [3] Y. Gao, W. Dong, C. Chen, J. Bu, W. Wu, and X. Liu, "ipath: Path inference in wireless sensor networks," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 517–528, 2014.
- [4] M. Laghate and D. Cabric, "Learning wireless networks' topologies using asymmetric granger causality," *IEEE J. Sel. Top. Signal Process.*, vol. 12, no. 1, pp. 233–247, 2017.
- [5] P. Sharma, D. J. Bucci, S. K. Brahma, and P. K. Varshney, "Communication network topology inference via transfer entropy," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 562–575, 2019.
- [6] P. Tilghman and D. Rosenbluth, "Inferring wireless communications links and network topology from externals using granger causality," in *Proc. MILCOM*. IEEE, 2013, pp. 1284–1289.
- [7] Z. Guo, V. M. McClelland, O. Simeone, K. R. Mills, and Z. Cvetkovic, "Multiscale wavelet transfer entropy with application to corticomuscular coupling analysis," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 2, pp. 771–782, 2021.
- [8] J. D. Finkle, J. J. Wu, and N. Bagheri, "Windowed granger causal inference strategy improves discovery of gene regulatory networks," *Proc. Nat. Acad. Sci.*, vol. 115, no. 9, pp. 2252–2257, 2018.
- [9] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muoz-Mar *et al.*, "Inferring causation from time series in earth system sciences," *Nature Commun.*, vol. 10, no. 1, pp. 1–13, 2019.
- [10] A. Bovet and H. A. Makse, "Influence of fake news in twitter during the 2016 us presidential election," *Nature Commun.*, vol. 10, no. 1, pp. 1–14, 2019.
- [11] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, pp. 424–438, 1969.
- [12] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, no. 2, p. 461, 2000.
- [13] M. Staniek and K. Lehnertz, "Symbolic transfer entropy," *Phys. Rev. Lett.*, vol. 100, no. 15, p. 158101, 2008.
- [14] J. Zhang, O. Simeone, Z. Cvetkovic, E. Abela, and M. Richardson, "Itene: Intrinsic transfer entropy neural estimator," *arXiv preprint arXiv:1912.07277*, 2019.
- [15] W. Liang, S. X. Ng, and L. Hanzo, "Cooperative overlay spectrum access in cognitive radio networks," *IEEE Commun. Surv. Tutorials*, vol. 19, no. 3, pp. 1924–1944, 2017.
- [16] Z. Liu, G. Ding, Z. Wang, S. Zheng, J. Sun, and Q. Wu, "Cooperative topology sensing of wireless networks with distributed sensors," *IEEE Trans. Cognit. Commun. Networking*, vol. 7, no. 2, pp. 524–540, 2020.
- [17] B. Deb, S. Bhatnagar, and B. Nath, "A topology discovery algorithm for sensor networks with applications to network management," 2002.
- [18] Y. Zeng, Y.-C. Liang, and R. Zhang, "Blindly combined energy detection for spectrum sensing in cognitive radio," *IEEE Signal Processing Lett.*, vol. 15, pp. 649–652, 2008.
- [19] C. Partridge, D. Cousins, A. W. Jackson, R. Krishnan, T. Saxena, and W. T. Strayer, "Using signal processing to analyze wireless data traffic," in *Proc. Workshop on Wirel. Secur.*, 2002, pp. 67–76.

- [20] M. G. Moore and M. A. Davenport, "Analysis of wireless networks using Hawkes processes," in *Proc. IEEE 17th Int. Workshop Signal Process. Adv. Wireless Commun.* IEEE, 2016, pp. 1–5.
- [21] H. Xu, M. Farajtabar, and H. Zha, "Learning granger causality for Hawkes processes," in *Proc. ICML*. PMLR, 2016, pp. 1717–1726.
- [22] E. Testi, E. Favarelli, L. Pucci, and A. Giorgetti, "Machine learning for wireless network topology inference," in *Proc. 13th Int. Conf. Signal Process. Commun. Syst. (ICSPCS)*. IEEE, 2019, pp. 1–7.
- [23] H. Elsegai, "Granger-causality inference in the presence of gaps: An equidistant missing-data problem for non-synchronous recorded time series data," *Physica A*, vol. 523, pp. 839–851, 2019.
- [24] E. Testi and A. Giorgetti, "Blind wireless network topology inference," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1109–1120, 2020.
- [25] A. K. Seth, "A matlab toolbox for granger causal connectivity analysis," *J. Neurosci. Methods*, vol. 186, no. 2, pp. 262–273, 2010.
- [26] J. Liu, S. Kumar, and D. P. Palomar, "Parameter estimation of heavy-tailed AR model with missing data via stochastic EM," *IEEE Trans. Signal Process.*, vol. 67, no. 8, pp. 2159–2172, 2019.
- [27] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- [28] O. Simeone *et al.*, "A brief introduction to machine learning for engineers," *Foundations and Trends® in Signal Processing*, vol. 12, no. 3-4, pp. 200–431, 2018.
- [29] M. Lacage and T. R. Henderson, "Yet another network simulator," in *ACM Int. Conf. Proc. Ser.*, 2006, pp. 12–es.
- [30] L. Campanile, M. Gribaudo, M. Iacono, F. Marulli, and M. Mastroianni, "Computer network simulation with ns-3: A systematic literature review," *Electronics*, vol. 9, no. 2, p. 272, 2020.