



King's Research Portal

DOI:

[10.1007/978-3-319-62389-4_9](https://doi.org/10.1007/978-3-319-62389-4_9)

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

Citation for published version (APA):

Charalampopoulos, P., Crochemore, M., Iliopoulos, C. S., Kociumaka, T., Pissis, S. P., Radoszewski, J., ... Walen, T. (2017). Efficient Enumeration of Non-Equivalent Squares in Partial Words with Few Holes. In Y. Cao, & J. Chen (Eds.), *Computing and Combinatorics: 23rd International Conference, COCOON 2017, Hong Kong, China, August 3-5, 2017, Proceedings* (Vol. 10392, pp. 99-111). Cham: Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-62389-4_9

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 providing details, and we will remove access to the work immediately and investigate your claim.

Efficient Enumeration of Non-Equivalent Squares in Partial Words with Few Holes

Panagiotis Charalampopoulos^{1,*}, Maxime Crochemore^{1,2},
Costas S. Iliopoulos¹, Tomasz Kociumaka^{3,**}, Solon P. Pissis¹,
Jakub Radoszewski^{1,3,***}, Wojciech Rytter^{3,***}, and Tomasz Walen^{3,***}

¹ Department of Informatics, King's College London, London, UK
[panagiotis.charalampopoulos,maxime.crochemore,
costas.ilopoulos,solon.pissis]@kcl.ac.uk

² Université Paris-Est, France

³ Faculty of Mathematics, Informatics and Mechanics,
University of Warsaw, Warsaw, Poland
[kociumaka,jrad,rytter,walen]@mimuw.edu.pl

Abstract. A word of the form WW for some word $W \in \Sigma^*$ is called a square, where Σ is an alphabet. A partial word is a word possibly containing holes (also called don't cares). The hole is a special symbol $\diamond \notin \Sigma$ which *matches* (agrees with) any symbol from $\Sigma \cup \{\diamond\}$. A *p-square* is a partial word matching at least one square WW without holes. Two p-squares are called *equivalent* if they match the same set of squares. We denote by $psquares(T)$ the number of non-equivalent p-squares which are factors of a partial word T . Let $PSQUARES_k(n)$ be the maximum value of $psquares(T)$ over all partial words of length n with at most k holes. We show asymptotically tight bounds:

$$c_1 \cdot \min(nk^2, n^2) \leq PSQUARES_k(n) \leq c_2 \cdot \min(nk^2, n^2)$$

for some constants $c_1, c_2 > 0$. We also present an algorithm that computes $psquares(T)$ in $\mathcal{O}(nk^3)$ time for a partial word T of length n with k holes. In particular, our algorithm runs in linear time for $k = \mathcal{O}(1)$ and its time complexity near-matches the maximum number of non-equivalent p-square factors in a partial word.

1 Introduction

A *word* is a sequence of letters from a given alphabet Σ . By Σ^* we denote the set of all words over Σ . A word of the form $U^2 = UU$, for some word U , is called a *square*. For a word W , a *square factor* is a factor of W which is a square. Enumeration of square factors in words is a well-studied topic, both from a combinatorial and from an algorithmic perspective. Obviously, a word W of

* Supported by the Graduate Teaching Scholarship scheme of the Department of Informatics at King's College London.

** Supported by Polish budget funds for science in 2013-2017 as a research project under the 'Diamond Grant' program.

*** Supported by the Polish National Science Center, grant no. 2014/13/B/ST6/00770.

length n may contain $\Theta(n^2)$ square factors (e.g. $W = a^n$), however, it is known that such a word contains only $\mathcal{O}(n)$ distinct square factors [14,17]; currently the best known upper bound is $\frac{11}{6}n$ [12]. Moreover, all distinct square factors of a word can be listed in $\mathcal{O}(n)$ time using the suffix tree [15] or the suffix array and the structure of runs (maximal repetitions) in the word [10].

A *partial word* is a sequence of letters from $\Sigma \cup \{\diamond\}$, where \diamond denotes a *hole*, that is, a don't care symbol. We assume that Σ is non-unary. Two symbols $a, b \in \Sigma \cup \{\diamond\}$ are said to *match* (denoted as $a \approx b$) if they are equal or one of them is a hole; note that this relation is not transitive. The relation of matching is extended in a natural way to partial words of the same length.

A partial word UV is called a *p-square* if $U \approx V$. Like in the context of words, a *p-square factor* of a partial word T is a factor being a p-square; see [2,7]. Alongside [2,6,7], we define a *solid square* (also called a *full square*) as a square of a word, and a *square subword* of a partial word T as a solid square that matches a factor of T .

We introduce the notion of *equivalence* of p-square factors in partial words. Let $sq\text{-val}(UV)$ denote the set of solid squares that match the partial word UV :

$$sq\text{-val}(UV) = \{WW : W \in \Sigma^*, WW \approx UV\}.$$

Example 1. $sq\text{-val}(a\diamond b a\diamond\diamond) = \{(aab)^2, (abb)^2\}$, with $\Sigma = \{a, b\}$.

Then p-squares UV and $U'V'$ are called *equivalent* if $sq\text{-val}(UV) = sq\text{-val}(U'V')$ (denoted as $UV \equiv U'V'$). For example, $a\diamond b a\diamond\diamond \equiv a\diamond\diamond \diamond\diamond b$, but $a\diamond b a\diamond\diamond \not\equiv a\diamond\diamond \diamond ab$.

Note that two p-square factors of a partial word T are equivalent in this sense if and only if they correspond to exactly the same set of square subwords. The number of non-equivalent p-square factors in a partial word T is denoted by $psquares(T)$. Our work is devoted to the enumeration of non-equivalent p-square factors in a partial word with a given number k of holes.

We say that $X^2 = XX$ is the *representative* (also called *general form*; see [6]) of a p-square UV , denoted as $repr(UV)$, if $XX \approx UV$ and $sq\text{-val}(XX) = sq\text{-val}(UV)$. (In other words, X is the “most general” partial word that matches both U and V .) It can be noted that the representative of a p-square is unique. Then $UV \equiv U'V'$ if and only if $repr(UV) = repr(U'V')$.

Example 2. $repr(a\diamond b a\diamond\diamond) = (a\diamond b)^2$, $repr(a\diamond\diamond \diamond ab) = (aab)^2$.

Previous studies on squares in partial words were mostly focused on combinatorics. They started with the case of $k = 1$ [6], in which case distinct square subwords correspond to non-equivalent p-square factors. It was shown that a partial word with one hole contains at most $\frac{7}{2}n$ distinct square subwords [4] ($3n$ for binary partial words [16]). Also a generalization of the three squares lemma (see [11]) was proposed for partial words [5]. As for a larger number of holes, the existing literature is devoted mainly to counting the number of distinct square subwords of a partial word [6,2] or all occurrences of p-square factors [3,2]. On the algorithmic side, [21] proved that the problem of counting distinct square subwords of a partial word is #P-complete and [13,20] and [7] showed quadratic- and nearly-quadratic-time algorithms for finding all occurrences of p-square factors and primitively-rooted p-square factors of a partial word, respectively.

Our combinatorial results. We prove that a partial word of length n with k holes contains $\mathcal{O}(nk^2)$ non-equivalent p-square factors. We also construct a family of partial words that contain $\Omega(nk^2)$ non-equivalent p-square factors, for $k = \mathcal{O}(\sqrt{n})$. This proves the aforementioned asymptotic bounds for $\text{PSQUARES}_k(n)$. Our work can be viewed as a generalization of the results on partial words with one hole [6,4,16] to $k \geq 1$ holes.

Our algorithmic results. We present an algorithm that reports all non-equivalent p-square factors in a partial word of length n with k holes in $\mathcal{O}(nk^3)$ time. In particular, our algorithm runs in linear time for $k = \mathcal{O}(1)$ and its time complexity near-matches the maximum number of non-equivalent p-square factors. We assume integer alphabet $\Sigma \subseteq \{1, \dots, n^{\mathcal{O}(1)}\}$. The main tool in the algorithm are two new types of non-standard runs in partial words and relations between them. We also use recently introduced advanced data structures from [18].

2 Preliminary Notation for Words and Partial Words

For a word $W \in \Sigma^*$, by $|W| = n$ we denote the length of W , and by $W[i]$, for $i = 1, \dots, n$, the i th letter of W . For $1 \leq i \leq j \leq n$, $W[i..j]$ denotes the *factor* of W equal to $W[i] \cdots W[j]$. A factor of the form $W[1..j]$ is called a *prefix*, a factor of the form $W[i..n]$ is called a *suffix*, and a factor that is both a prefix and a suffix of W is called a *border* of W . A positive integer q is called a *period* of W if $W[i] = W[i+q]$ for all $i = 1, \dots, n-q$. In this case, $W[1..q]$ is called a *string period* of W . W has a period q if and only if it has a border of length $n-q$; see [8]. Two equal-length words V and W are called *cyclic shifts* if there are words X, Y such that $V = XY$ and $W = YX$. A word W is called *primitive* if there is no word U and integer $k > 1$ such that $U^k = W$. Note that the shortest string period of W is always primitive. Every primitive word W has the following *synchronization property*: W is not equal to any of its non-trivial cyclic shifts [8].

For a partial word T we use the same notation as for words: $|T| = n$ for its length, $T[i]$ for the i th letter, $T[i..j]$ for a factor. If T does not contain holes, then it is called *solid*. The relation of matching on $\Sigma \cup \{\diamond\}$ is defined as: $a \approx a$, $\diamond \approx a$, and $a \approx \diamond$ for all $a \in \Sigma \cup \{\diamond\}$. We define an operation \wedge such that $a \wedge a = a \wedge \diamond = \diamond \wedge a = a$ for all $a \in \Sigma \cup \{\diamond\}$, and otherwise $a \wedge b$ is undefined. Two equal-length partial words T and S are said to *match* (denoted as $T \approx S$) if $T[i] \approx S[i]$ for all $i = 1, \dots, n$. In this case, by $S \wedge T$ we denote the partial word $S[1] \wedge T[1], \dots, S[n] \wedge T[n]$. If $U \approx T[i..i+|U|-1]$ for a partial word U , then we say that U occurs in T at position i . Also note that if UV is a p-square, then $\text{repr}(UV) = (U \wedge V)^2$. A *quantum period* of T is a positive integer q such that $T[i] \approx T[i+q]$ for all $i = 1, \dots, n-q$. A *deterministic period* of T is an integer q such that there exists a word W such that $W \approx T$ and W has a period q . T is called *quantum (deterministically) periodic* if it has a quantum (deterministic) period q such that $2q \leq n$.

An integer j is an *ambiguous length* in the partial word T if there are two holes in T at distance $j/2$. A p-square is called *ambiguous* if its representative

is non-solid. Note that if a p-square factor in T is ambiguous, then the p-square has an ambiguous length (the converse is not always true). The p-square factors of T of non-ambiguous length have solid representatives.

Example 3. Let $T = ab\diamond\diamond ba\diamond aaba\diamond b$. For T , 4 is a non-ambiguous length. T contains four non-equivalent classes of p-squares of length 4: $a\diamond aa$ with representative $(aa)^2$, $ab\diamond\diamond \equiv \diamond ba\diamond \equiv aba\diamond$ with representative $(ab)^2$, $\diamond\diamond ba \equiv ba\diamond a$ with representative $(ba)^2$, and $b\diamond\diamond b$ with representative $(bb)^2$. On the other hand, 6 is an ambiguous length in T . T contains four non-equivalent classes of p-squares of length 6: $aaba\diamond b$ with representative $(aab)^2$, $ab\diamond\diamond ba \equiv a\diamond aaba$ with representative $(aba)^2$, $\diamond aaba\diamond$ with representative $(baa)^2$, and $b\diamond\diamond ba\diamond$ with representative $(ba\diamond)^2$. Note that only the last one is an ambiguous p-square. Overall, T contains 14 non-equivalent p-squares.

3 Combinatorial Bounds

3.1 Lower Bound

We say that a set A of positive integers is an (m, t) -cover if the following conditions hold:

- (1) For each $d \geq m$, A contains at most one pair of elements with difference d ;
- (2) $|\{j - i \geq m : i, j \in A\}| \geq t$.

For a set $A \subseteq \{1, \dots, n\}$ we denote by $w_{A,n}$ the partial word of length n over the alphabet Σ such that $w_{A,n}[i] = \diamond \Leftrightarrow i \in A$, and $w_{A,n}[i] = a$ otherwise.

Lemma 4. *Assume that $A \subseteq \{1, \dots, n\}$ is an (m, t) -cover such that $m = \Theta(n)$, $|A| = k$, and $t = \Omega(k^2)$. Let $\Sigma = \{a, b\}$ be the alphabet. Then*

$$\text{psquares}(a^{n-2} \cdot w_{A,n} \cdot a^{n-2}) = \Omega(n \cdot k^2).$$

Proof. Each even-length factor of $a^{n-2} \cdot w_{A,n} \cdot a^{n-2}$ is a p-square. Let \mathcal{Z} be the set of these factors X which contain two positions i, j containing holes with $|j - i| \geq m$ and $|X| = 2|j - i|$. As A is an (m, t) -cover, i and j are determined uniquely by $d = |j - i|$. Then all elements of \mathcal{Z} are pairwise non-equivalent p-squares. The size of \mathcal{Z} is $\Omega(nt)$ which is $\Omega(n \cdot k^2)$. This completes the proof. \square

Example 5. Let $n = 5$, $m = 4$, and $t = 1$. $aaa\diamond aaa\diamond aaa$ has 4 non-equivalent p-square factors of length 8 if $\Sigma = \{a, b\}$. If $\Sigma = \{a\}$, all of them are equivalent.

Theorem 6. *For every positive integer n and $k \leq \sqrt{2n}$, there is a partial word of length n with k holes that contains $\Omega(nk^2)$ non-equivalent p-square factors.*

Proof. Due to Lemma 4, it is enough to construct a suitable set A . By monotonicity, we may assume that k and n are even. We take:

$$A = \{1, \dots, \frac{k}{2}\} \cup \{j \cdot \frac{k}{2} + \frac{n}{2} : 1 \leq j \leq \frac{k}{2}\}.$$

We claim that A is an $(\frac{n}{2}, \frac{k^2}{4})$ -cover for $t = \Omega(k^2)$. Indeed, take any $i \in \{1, \dots, \frac{k}{2}\}$ and j satisfying the above condition. Then $j \cdot \frac{k}{2} + \frac{n}{2} - i \geq \frac{n}{2}$ and all such values are distinct; hence, $t = \frac{k^2}{4}$. The thesis follows from the claim. \square

3.2 Upper Bound

Let T be a partial word of length n with k holes. The proof of the upper bound for ambiguous lengths is easy.

Lemma 7. *There are at most nk^2 p -square factors of ambiguous length in T .*

Proof. The number of ambiguous lengths is at most $\binom{k}{2}$, since we have $\binom{k}{2}$ possible distances between k holes. Consequently, the number of p -squares with such lengths is at most nk^2 . \square

Each of the remaining p -square factors of T has a solid representative. We say that a solid square W^2 has a *solid occurrence* in T if T contains a factor equal to W^2 . By the following fact, there are at most $2n$ non-equivalent p -square factors of T with solid occurrences.

Fact 8 ([14,17,12]). *Every position of a (solid) word contains at most two rightmost occurrences of squares.*

We say that a solid square is a *u-square* in T if it occurs in T , does not have a solid occurrence in T , and has a non-ambiguous length. We denote by \mathcal{U} the set of u -squares for T .

Observation 9. *Each u -square in T corresponds in a one-to-one way to an equivalence class of p -square factors of T which have non-ambiguous length and do not have a solid occurrence in T .*

Thus it suffices to bound $|\mathcal{U}|$. This is the essential part of the proof.

Let $\alpha = \frac{1}{2k+2}$ and

$$\mathcal{U}(\ell) = \{W^2 \in \mathcal{U} : 2\ell \leq |W|^2 \leq 2(\ell + \lfloor \ell\alpha \rfloor)\}.$$

Also denote by $\mathcal{U}_i(\ell)$ (and $\mathcal{U}_{last_i}(\ell)$) the set of words of $\mathcal{U}(\ell)$ which have an occurrence (the last occurrence, respectively) at position i in T . The next lemma follows from the pigeonhole principle and periodicity of (solid) words.

Lemma 10. *Suppose that $\ell \geq \frac{1}{\alpha}$ and $|\mathcal{U}_i(\ell)| \geq 2$. Let $\Delta = \lfloor \ell\alpha \rfloor$. There exist positions s, s' such that:*

- $s \in [i, i + \ell - 2\Delta]$,
- $s' \in [s + \ell, s + \ell + \Delta]$,
- $T[s..s + 2\Delta - 1] = T[s'..s' + 2\Delta - 1]$ is solid and periodic.

Proof. Let $T[i..i + 2d - 1]$ be a u -square from $\mathcal{U}_i(\ell)$. Consider positions $x_j = i + 2j\Delta$ and $y_j = x_j + d$ for $0 \leq j \leq k$. Note that factors $X_j = T[x_j..x_j + 2\Delta - 1]$ and $Y_j = T[y_j..y_j + 2\Delta - 1]$ match; see Fig. 1. Moreover, factors X_0, \dots, X_k and Y_0, \dots, Y_k are disjoint because $2(k+1)\Delta \leq 2(k+1)\frac{\ell}{2k+2} = \ell$. By the pigeonhole principle, we can choose j so that X_j and Y_j are solid, i.e., $X_j = Y_j$. We set $s = x_j$ and $s' = y_j$.

It remains to prove that $X_j = Y_j$ is periodic. Let $T[i..i + 2d' - 1]$ (with $d' \neq d$) be another u -square in $\mathcal{U}_i(\ell)$, and let $Y'_j = T[x_j + d'..x_j + d' + 2\Delta - 1]$. Note that $Y'_j \approx X_j = Y_j$ and factors Y_j and Y'_j have an overlap of $2\Delta - |d - d'|$ positions being a border of Y_j . Consequently, $|d - d'| \leq \lfloor \ell\alpha \rfloor = \Delta$ is a period of Y_j . \square

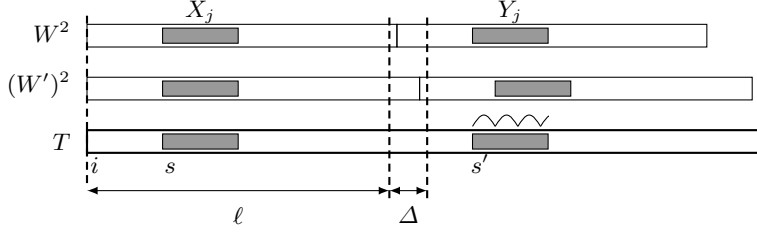


Fig. 1. Situation from the proof of Lemma 10; $W^2, (W')^2 \in \mathcal{U}_i(\ell)$. Occurrences of $X_j = Y_j$ are denoted by dark rectangles. $\mathcal{U}_i(\ell)$ is the set of all u-squares having an occurrence in i with center in the window of size Δ .

We denote by $\mathcal{I}_{i,\ell}$ the interval $[i, i + 2(\ell + \lfloor \ell\alpha \rfloor) - 1]$. Let $\#\diamond([a, b])$ denote the number of holes in $T[a..b]$. Our upper bound for partial words is based on the following key lemma; it is a property of partial words similar to Fact 8.

Lemma 11. $|\mathcal{U}last_i(\ell)| = \mathcal{O}(\#\diamond(\mathcal{I}_{i,\ell}))$.

Proof. Denote $k' = \#\diamond(\mathcal{I}_{i,\ell})$. If $k' = 0$, then $|\mathcal{U}last_i(\ell)| = 0$. From now we assume that $k' \geq 1$. Assume that $|\mathcal{U}last_i(\ell)| \geq 2$. Let p be the shortest period of the equal periodic factors $X = T[s..s + 2\Delta - 1]$ and $Y = T[s'..s' + 2\Delta - 1]$ from the previous lemma. We consider three types of u-squares $W^2 \in \mathcal{U}last_i(\ell)$:

- Type (a):** W^2 has period p ;
- Type (b):** W has period p but W^2 does not have period p ;
- Type (c):** W does not have period p .

At most 1 u-square of type (a). Observe that the length of W is a multiple of its shortest period p (this is due to the synchronization property for the string period of W). Consequently, if we have two u-squares of type (a) occurring at position i and with the same shortest period p , then the shorter u-square also occurs at position $i + p$. This contradicts the definition of $\mathcal{U}last_i(\ell)$.

At most $k' + 1$ u-squares of type (b). Suppose to the contrary that there are at least $k' + 2$ u-squares of type (b), of lengths $d_1 < \dots < d_{k'+2}$. Note that $Y'_j := T[s + d_j..s + d_j + 2\Delta - 1]$ matches $X = Y$ due to a u-square of length $2d_j$. Moreover, the factors Y and Y'_j have an overlap of at least $\Delta \geq p$ positions, so the string periods of Y'_j and Y must be synchronized. Consequently, the values $d_j \bmod p$ are all the same (and non-zero, as these are not squares of type (a)).

Consider the shortest W^2 and the longest $(W')^2$ of these u-squares and the factor $Z = T[i + d_1..i + d_{k'+2} - 1]$. It matches a prefix P of length $d_{k'+2} - d_1$ of W and a suffix S of the same length of W' . Both P and S have period p ; however, their string periods of length p are not equal (again, due to synchronization property), as p does not divide d_1 . Consequently, in every factor of length p in Z there must be a hole. This yields $\lfloor |Z|/p \rfloor = (d_{k'+2} - d_1)/p \geq k' + 1$ holes in total, a contradiction.

At most $4k' + 2$ u-squares of type (c). Let $d = |W|$. Let us extend the occurrence of X in W at position $s - i + 1$ to a maximal factor $W[j'..j]$ with period p . Note that $j' > 1$ or $j < d$ as W^2 is not of type (b). Below, we assume $j < d$; the other subcase is handled in an analogous way. Consider the positions $j_1 = i + j$ and $j_2 = i + d + j$ of T . We will show that there are at most $2k' + 1$ possible pairs (j_1, j_2) across the u-squares $W \in \mathcal{U}last_i(\ell)$, i.e., at most $2k' + 1$ corresponding u-squares, as $d = j_2 - j_1$.

Positions $T[j_1]$ and $T[j_2]$ cannot both contain holes, as $2d$ is a non-ambiguous length. If $T[j_1]$ is not a hole, then it is determined uniquely as the first position where the deterministic period p breaks, starting from the position s , i.e., j_1 is the smallest index such that $T[s..j_1]$ does not have deterministic period p . The same holds for j_2 and s' ; this is also due to the fact that Y and the occurrence of X at position $s + d$ have an overlap of at least $\Delta \geq p$ positions, so they are synchronized. Hence, if neither $T[j_1]$ nor $T[j_2]$ is a hole, then (j_1, j_2) is determined uniquely. Otherwise, if $T[j_1]$ or $T[j_2]$ is a hole, then the other position is determined uniquely, so there are at most $2k'$ choices. This concludes the proof. \square

Theorem 12. *The number of non-equivalent p -square factors in a partial word T of length n with k holes is $\mathcal{O}(\min(nk^2, n^2))$.*

Proof. The $\mathcal{O}(n^2)$ bound is obvious. Due to Lemma 7 there are at most nk^2 p-squares of ambiguous length in T . Let us consider p-squares of non-ambiguous lengths. By Fact 8, among them there are $\mathcal{O}(n)$ non-equivalent p-squares with a solid occurrence. From now on we count only non-equivalent non-ambiguous p-squares without a solid occurrence, i.e., different u-squares.

Clearly, there are $\mathcal{O}(nk)$ different u-squares of length smaller than $\frac{2}{\alpha}$. Let $\ell \geq \frac{1}{\alpha}$ and $r = 2(\ell + \lfloor \ell\alpha \rfloor)$. By Lemma 11:

$$|\mathcal{U}(\ell)| = \sum_{i=1}^n |\mathcal{U}last_i(\ell)| = \mathcal{O}\left(\sum_{i=1}^n \#\diamond(\mathcal{I}_{i,\ell})\right) = \mathcal{O}(k\ell). \quad (1)$$

The last equality is based on the fact that each of the k holes in T is counted in at most $2r$ terms $\#\diamond(\mathcal{I}_{i,\ell})$.

Let us consider a family of endpoints $r_j = \left\lceil \frac{n}{(1+\alpha)^j} \right\rceil$ for $j \geq 0$ and let $t = \max\{j : r_j > 1\}$. One can check that $\mathcal{U} = \bigcup_{j=0}^t \mathcal{U}(r_{j+1})$.

By (1), the total number of u-squares of length at least $\frac{2}{\alpha}$ in T is at most:

$$\begin{aligned} \sum_{j=1}^{t+1} |\mathcal{U}(r_j)| &= \mathcal{O}\left(\sum_{j=1}^{t+1} kr_j\right) = \mathcal{O}\left(k \sum_{j=1}^{t+1} \left(1 + \frac{n}{(1+\alpha)^j}\right)\right) \\ &= \mathcal{O}\left(k \log_{1+\alpha} n + \sum_{j=0}^{\infty} \frac{nk}{(1+\alpha)^j}\right) = \mathcal{O}\left(\frac{k \log n}{\alpha} + \frac{nk}{1 - \frac{1}{1+\alpha}}\right) = \mathcal{O}(nk^2). \quad \square \end{aligned}$$

4 Runs Toolbox for Partial Words

A *run* (also called a maximal repetition) in a word W is a triple (a, b, q) such that $W[a..b]$ is periodic with period q ($2q \leq b - a + 1$) and the interval $[a, b]$ cannot be extended to the left nor to the right without violating the above property, that is, $W[a - 1] \neq W[a + q - 1]$ and $W[b - q + 1] \neq W[b + 1]$, provided that the respective positions exist. The *exponent* of a run is defined as $\frac{b-a+1}{q}$. A word of length n has $\mathcal{O}(n)$ runs and they can all be computed in $\mathcal{O}(n)$ time [19,1].

From a run (a, b, q) we can produce all triples (a, b, kq) for integer $k \geq 1$ such that $2kq \leq b - a + 1$; we call such triples *generalized runs*. That is, the period of a generalized run need not be the shortest period. The number of generalized runs is also $\mathcal{O}(n)$ as the sum of exponents of runs is $\mathcal{O}(n)$ [19,1].

For a partial word T , we call a triple (a, b, q) a *quantum generalized run* (Q-run, for short) in T if $T[a..b]$ is quantum periodic with period q and none of the partial words $T[a - 1..b]$ and $T[a..b + 1]$ (if it exists) has the quantum period q ; for an example see Fig. 2.

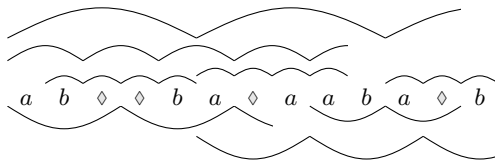


Fig. 2. A partial word together with all its Q-runs.

Generalized runs in words are strongly related to squares: (1) a square of length $2q$ belongs to a generalized run of period q and, moreover, (2) all factors of length $2q$ of a generalized run with period q are squares being each other's cyclic shifts. Unfortunately, Q-runs in partial words have only property (1). However, we introduce a type of run in partial words that has a property analogous to (2). A *pseudorun* is a triple (a, b, q) such that:

- (a) $T[a..b]$ is quantum periodic with period q
- (b) $T[i - q] \wedge T[i] = T[i] \wedge T[i + q]$ for all i such that $i - q, i + q \in [a, b]$,
- (c) none of the partial words $T[a - 1..b]$ and $T[a..b + 1]$ (if exists) satisfies the conditions (a) and (b).

We say that a p-square factor $T[c..d]$ is *induced* by the pseudorun (a, b, q) if $d - c + 1 = 2q$ and $[c, d] \subseteq [a, b]$.

Example 13. The partial word from Fig. 2 contains two Q-runs with period 2: $(1, 9, 2)$ that corresponds to factor $ab\diamond\diamondbaaa$ and $(9, 12, 2)$ that corresponds to factor $ba\diamond$. The partial word contains five pseudoruns with this period: $(1, 4, 2)$: $ab\diamond$, $(2, 5, 2)$: $b\diamond b$, $(3, 8, 2)$: $\diamond\diamondba\diamond a$, $(6, 9, 2)$: $a\diamond a a$, and $(9, 12, 2)$: $ba\diamond$. All but one of these pseudoruns induce exactly one p-square; the pseudorun $(3, 8, 2)$ induces two non-equivalent p-squares: $\diamond\diamondba$ and $\diamond ba\diamond$.

Observation 14. (1) Every p -square factor in T is induced by a pseudorun.
(2) All factors of length $2q$ of a pseudorun with period q are p -squares and their representatives are each other's cyclic shifts.

5 The Algorithm

We design an $\mathcal{O}(nk^3)$ -time algorithm for enumerating non-equivalent p -squares in a partial word T of length n with k holes. We assume that Σ is an ordered integer alphabet and that \diamond is smaller than all the letters from Σ . Then any two factors of T can be lexicographically compared using the suffix array of T in $\mathcal{O}(1)$ time after $\mathcal{O}(n)$ -time preprocessing [8]. The first two steps of the algorithm are computing all Q-runs in T and decomposing Q-runs into pseudoruns. The final phase consists in grouping pseudoruns in T by the representatives of induced p -squares, which lets us enumerate non-equivalent p -squares.

5.1 Computing Q-runs

We classify Q-runs into *solid Q-runs* that do not contain a hole and the remaining *non-solid Q-runs*. A solid Q-run is a generalized run in a maximal solid factor of T that is not adjacent to a hole in T . Thus all solid Q-runs can be computed in $\mathcal{O}(n)$ time using any linear-time algorithm for computing runs in words [19,1].

The length of the *longest common compatible prefix* of two positions i, j , denoted $lccp(i, j)$, is the largest ℓ such that $T[i..i + \ell - 1] \approx T[j..j + \ell - 1]$. Symmetrically, we can define $lccs(i, j)$ as the length of the longest common compatible suffix of $T[1..i]$ and $T[1..j]$. After $\mathcal{O}(nk)$ -time preprocessing, queries for $lccp$ (hence, queries for $lccs$) can be answered on-line in $\mathcal{O}(1)$ time [9].

For every position i containing a hole and integer $q \in \{1, \dots, n\}$, we can use the $lccp$ - and $lccs$ -queries to check if there is a Q-run with period q containing the position i . If the Q-run is to contain i anywhere except for its last q positions, we can compute $a = i - lccs(i, i + q) + 1$, $b = i + q + lccp(i, i + q) - 1$ and check if $b - a + 1 \geq 2q$; if so, the sought Q-run is (a, b, q) . A symmetric test with $i - q$ and i can be used to check for a Q-run containing i among its last q positions.

Clearly, this procedure works in $\mathcal{O}(nk)$ time. Therefore, the number of Q-runs is at most $\mathcal{O}(nk)$. The same Q-run may be reported several times; therefore, in the end we remove repeating triples (a, b, q) via radix sort. Together with the $\mathcal{O}(n)$ -time computation of solid Q-runs we arrive at the following lemma.

Lemma 15. *A partial word of length n with k holes contains $\mathcal{O}(nk)$ Q-runs and they can all be computed in $\mathcal{O}(nk)$ time.*

5.2 Computing Pseudoruns

Q-runs correspond to maximal factors of T that satisfy only the condition (a) of a pseudorun. Hence, every pseudorun is a factor of a Q-run.

A position i inside a Q-run $\beta = (a, b, q)$ is called a *break point* if $a \leq i - q < i + q \leq b$ and $T[i - q] \wedge T[i] \neq T[i] \wedge T[i + q]$.

Observation 16. i is a break point for (a, b, q) if and only if $a \leq i - q < i + q \leq b$, $T[i] = \diamond$, and $T[i - q] \neq T[i + q]$.

By $\Gamma(\beta)$ we denote the set of all break points of a Q-run β . The Q-run can be decomposed into $|\Gamma(\beta)| + 1$ pseudoruns: if i is the first break point in β , then we have a pseudorun $(a, i + q - 1, q)$ and continue the decomposition for $(i - q + 1, b, q)$. Consecutive pseudoruns in the decomposition overlap by $2q - 1$ positions. See Fig. 3 for an abstract illustration.

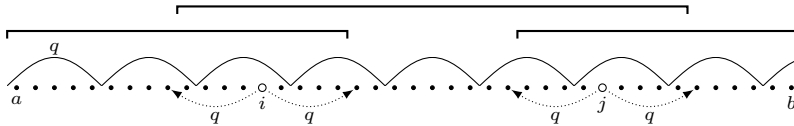


Fig. 3. A Q-run (a, b, q) with break points at positions i and j is decomposed into three pseudoruns: $(a, i + q - 1, q)$, $(i - q + 1, j + q - 1, q)$, and $(j - q + 1, b, q)$.

Lemma 17. $\sum_{\beta \in Q\text{-runs}(T)} |\Gamma(\beta)| \leq nk$.

Proof. Consider all Q-runs β of period q . Every two overlap by at most $q - 1$ positions, so the $\Gamma(\beta)$ sets are pairwise disjoint and their sizes sum up to at most k . Summing up over all $q = 1, \dots, n/2$, we arrive at the conclusion. \square

Lemma 17 shows that there are $\mathcal{O}(nk)$ pseudoruns (we use the fact that, by Lemma 15, there are $\mathcal{O}(nk)$ Q-runs). They can be computed in $\mathcal{O}(nk^2)$ time by inspecting all the holes inside each Q-run β and checking which of them are break points in β .

Lemma 18. A partial word of length n with k holes contains $\mathcal{O}(nk)$ pseudoruns and they can all be computed in $\mathcal{O}(nk^2)$ time.

5.3 Grouping Pseudoruns and Reporting Squares

We define the *representative* of a pseudorun $\beta = (a, b, q)$ as

$$\text{repr}(\beta) = \text{lex-min}\{\text{repr}(T[i..i + 2q - 1]) : a \leq i \leq b - 2q + 1\}.$$

First, let us show how to group pseudoruns by equal representatives. This part of our algorithm builds upon the methods for grouping runs in words from [10].

We use a separate approach for solid and for non-solid pseudoruns. Each solid pseudorun corresponds to a solid Q-run. Hence, there are $\mathcal{O}(n)$ of them and they can all be grouped using the approach of [10] in $\mathcal{O}(n)$ time.

We say that a partial word U is a d -fragment of T if U is a factor of T with symbols at d positions substituted with other symbols. Obviously, a d -fragment can be represented in $\mathcal{O}(d)$ space. The following lemma is a consequence of Observation 18 from [18] and Theorem 23 from [18].

Lemma 19 ([18]). *For a word of length n , after $\mathcal{O}(n)$ -time preprocessing:*

- (a) *Any two d -fragments can be compared lexicographically in $\mathcal{O}(d)$ time;*
- (b) *The minimal cyclic shift of a d -fragment can be computed in $\mathcal{O}(d^2)$ time.*

Lemma 20. *After $\mathcal{O}(n)$ -time preprocessing, for any pseudorun β , $\text{repr}(\beta)$ represented as a k -fragment can be computed in $\mathcal{O}(k^2)$ time.*

Proof. Let $\beta = (a, b, q)$. Knowing the positions of holes in T , we can represent $\text{repr}(T[a..a + 2q - 1]) = U^2$ as a k -fragment (the positions with holes of the p-square are filled with single symbols). By Lemma 19(b), we can find the minimal cyclic shift of the k -fragment in $\mathcal{O}(k^2)$ time. The cyclic shift can be represented as a k -fragment as well. We apply this to find $(U')^2$, the minimal cyclic shift of U^2 . Then $\text{repr}(\beta) = (U')^2$. \square

We group non-solid pseudoruns by their periods first; let \mathcal{R}_q be the set of non-solid pseudoruns with period q . From what we have already observed, we see that every pseudorun from \mathcal{R}_q can overlap with at most six other pseudoruns from \mathcal{R}_q : two that come from the same Q-run and two that come from each of the neighbouring Q-runs with period q . Hence, each hole position is contained in at most seven pseudoruns from \mathcal{R}_q , and $|\mathcal{R}_q| \leq 7k$. The representatives of pseudoruns from \mathcal{R}_q can be sorted using $\mathcal{O}(k)$ -time comparison (Lemma 19(a)). Thus the time complexity for sorting and grouping all pseudoruns from \mathcal{R}_q is $\mathcal{O}(k^2 \log k)$, which gives $\mathcal{O}(nk^2 \log k)$ in total.

By Observation 14, the representatives of all p-squares induced by a pseudorun β are cyclic shifts of $\text{repr}(\beta)$. Thus only pseudoruns from the same group may induce equivalent p-squares. For each pseudorun β we can specify an interval $I(\beta)$ of cyclic shift values of induced p-squares. Then all non-equivalent p-squares induced by pseudoruns in the same group can be reported by carefully processing the intervals $I(\beta)$ as in [10]. This processing takes time linear in the number of all intervals from all groups and n , i.e., $\mathcal{O}(nk)$ time. This concludes the algorithm.

Theorem 21. *All non-equivalent p-squares in a partial word of length n with k holes can be reported (as factors of the partial word) in $\mathcal{O}(nk^3)$ time.*

Proof. Lemma 18 shows that there are $\mathcal{O}(nk)$ pseudoruns in a partial word and they can all be computed in $\mathcal{O}(nk^2)$ time. Solid pseudoruns can be handled separately in $\mathcal{O}(n)$ time. Lemma 20 lets us find the representatives of non-solid pseudoruns in $\mathcal{O}(nk^3)$ time. In the end, we group those pseudoruns by the representatives in $\mathcal{O}(nk^2 \log k)$ time and use the approach from [10] to report all non-equivalent p-squares induced by each group in $\mathcal{O}(nk)$ time. \square

References

1. Bannai, H., I, T., Inenaga, S., Nakashima, Y., Takeda, M., Tsuruta, K.: A new characterization of maximal repetitions by Lyndon trees. In: Indyk, P. (ed.) 26th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015. pp. 562–571. SIAM (2015)

2. Blanchet-Sadri, F., Bodnar, M., Nikkel, J., Quigley, J.D., Zhang, X.: Squares and primitivity in partial words. *Discrete Applied Mathematics* 185, 26–37 (2015)
3. Blanchet-Sadri, F., Jiao, Y., Machacek, J.M., Quigley, J., Zhang, X.: Squares in partial words. *Theoretical Computer Science* 530, 42–57 (2014)
4. Blanchet-Sadri, F., Mercaş, R.: A note on the number of squares in a partial word with one hole. *Informatique Théorique et Applications* 43(4), 767–774 (2009)
5. Blanchet-Sadri, F., Mercaş, R.: The three-squares lemma for partial words with one hole. *Theoretical Computer Science* 428, 1–9 (2012)
6. Blanchet-Sadri, F., Mercaş, R., Scott, G.: Counting distinct squares in partial words. *Acta Cybernetica* 19(2), 465–477 (2009)
7. Blanchet-Sadri, F., Nikkel, J., Quigley, J.D., Zhang, X.: Computing primitively-rooted squares and runs in partial words. In: Kratochvíl, J., Miller, M., Fronček, D. (eds.) *Combinatorial Algorithms, IWOCA 2014*. LNCS, vol. 8986, pp. 86–97. Springer (2014)
8. Crochemore, M., Hancart, C., Lecroq, T.: *Algorithms on Strings*. Cambridge University Press (2007)
9. Crochemore, M., Iliopoulos, C.S., Kociumaka, T., Kubica, M., Langiu, A., Radoszewski, J., Rytter, W., Szreder, B., Waleń, T.: A note on the longest common compatible prefix problem for partial words. *Journal of Discrete Algorithms* 34, 49–53 (2015)
10. Crochemore, M., Iliopoulos, C.S., Kubica, M., Radoszewski, J., Rytter, W., Waleń, T.: Extracting powers and periods in a word from its runs structure. *Theoretical Computer Science* 521, 29–41 (2014)
11. Crochemore, M., Rytter, W.: Squares, cubes, and time-space efficient string searching. *Algorithmica* 13(5), 405–425 (1995)
12. Deza, A., Franek, F., Thierry, A.: How many double squares can a string contain? *Discrete Applied Mathematics* 180, 52–69 (2015)
13. Diaconu, A., Manea, F., Tiseanu, C.: Combinatorial queries and updates on partial words. In: Kutylowski, M., Charatonik, W., Gębala, M. (eds.) *Fundamentals of Computation Theory, FCT 2009*. LNCS, vol. 5699, pp. 96–108. Springer (2009)
14. Fraenkel, A.S., Simpson, J.: How many squares can a string contain? *Journal of Combinatorial Theory. Series A* 82(1), 112–120 (1998)
15. Gusfield, D., Stoye, J.: Linear time algorithms for finding and representing all the tandem repeats in a string. *Journal of Computer and System Sciences* 69(4), 525–546 (2004)
16. Halava, V., Harju, T., Kärki, T.: On the number of squares in partial words. *RAIRO – Theoretical Informatics and Applications* 44(1), 125–138 (2010)
17. Ilie, L.: A simple proof that a word of length n has at most $2n$ distinct squares. *Journal of Combinatorial Theory. Series A* 112(1), 163–164 (2005)
18. Kociumaka, T.: Minimal suffix and rotation of a substring in optimal time. In: Grossi, R., Lewenstein, M. (eds.) *Combinatorial Pattern Matching, CPM 2016*. LIPIcs, vol. 54, pp. 28:1–28:12. Schloss Dagstuhl (2016)
19. Kolpakov, R.M., Kucherov, G.: Finding maximal repetitions in a word in linear time. In: *40th Annual Symposium on Foundations of Computer Science, FOCS 1999*. pp. 596–604. IEEE Computer Society (1999)
20. Manea, F., Mercaş, R., Tiseanu, C.: An algorithmic toolbox for periodic partial words. *Discrete Applied Mathematics* 179, 174–192 (2014)
21. Manea, F., Tiseanu, C.: Hard counting problems for partial words. In: Dediu, A., Fernau, H., Martín-Vide, C. (eds.) *Language and Automata Theory and Applications, LATA 2010*. LNCS, vol. 6031, pp. 426–438. Springer (2010)