Vesa Halava | Tero Harju | Tomi Kärki

On the number of squares in partial words

Turku Centre for Computer Science

TUCS Technical Report
No 896, February 2009

On the number of squares in partial words

Vesa Halava
Department of Mathematics and
TUCS - Turku Centre for Computer Science
University of Turku, FI-20014 Turku, Finland
vehalava@utu.fi
Tero Harju
Department of Mathematics and
TUCS - Turku Centre for Computer Science
University of Turku, FI-20014 Turku, Finland
harju@utu.fi
Tomi Kärki
Département de Mathématiques
Université de Liège
Grande Traverse 12, B-4000 Liège, Belgium
topeka@utu.fi

TUCS Technical Report
No 896, February 2009

Abstract

The theorem of Fraenkel and Simpson states that the maximum number of distinct squares that a word w of length n can contain is less than $2n$. This is based on the fact that no more than two squares can have their last occurrences starting at the same position. In this paper we show that the maximum number of the last occurrences of squares per position in a partial word containing one hole is $2k$, where k is the size of the alphabet. Moreover, we prove that the number of distinct squares in a partial word with one hole and of length n is less than $4n$, regardless of the size of the alphabet. For binary partial words, this upper bound can be reduced to $3n$.

Keywords: Squares, partial words, theorem of Fraenkel and Simpson

TUCS Laboratory
Discrete Mathematics for Information Technology

1 Introduction

In combinatorics on words, factors of the form ww, i.e., squares can be studied from two perspectives. On one hand, one may try to avoid squares by constructing square-free words. A classical example of an infinite square-free word over a 3-letter alphabet is obtained from the famous Thue-Morse word [1] using a certain mapping; see [14, 15]. On the other hand, one may try to maximize the number of square factors in a word. The theorem of Fraenkel and Simpson states that a word of length n contains always less than $2n$ distinct squares [6]. A very short proof for this and an improved upper bound $2n - \Theta(\log n)$ was given by Ilie in [9] and [10]. However, based on the numerical evidence, the conjectured bound is n.

In this paper we consider squares in partial words, which are words with "do not know"-symbols $\diamond$ called holes. Here a square is a factor of the form ww', where w and w' are compatible. Compatibility means that words of the same length agree on each position which does not contain a hole. Partial words were first introduced by Berstel and Boasson in [2]. The theory of partial words has developed rapidly in recent years and many classical topics in combinatorics on words have been revisited for this generalization; see [3]. For example, the present authors proved in [8] that there exist uncountably many infinite square-free partial words over a 3-letter alphabet containing infinitely many holes. Note that for square-freeness, we must allow unavoidable squares $a\diamond$ and $\diamond a$ for (some) letters a. For other results on repetition-freeness in partial words, see [7] and [13].

In this paper our aim is to generalize the theorem of Fraenkel and Simpson for partial words containing one hole. This problem was already investigated by Blanchet-Sadri *et al.* in [4]. They proved that the number of distinct full words u^2 compatible with factors in a partial word with h holes and of length n increases polynomially with respect to k, where $k \geq 2$ is the size of the alphabet. Moreover, they showed that, for partial words with one hole, there may be more than two squares that have their last compatible occurrences starting at the same position. They also gave an intricate proof for the statement that in the above described case the hole must be in the shortest square.

A partial word containing one hole and $k+1$ squares whose last compatible occurrences start at the first position was given in [4]. In Section 3 we improve this example by constructing a word with $2k$ last compatible occurrences of squares starting at position one. We also show that this bound $2k$ is maximal. In Section 4 we prove that if a position contains three last occurrences of squares, then the longest square must be twice as long as the shortest square. As a corollary, we get a short proof for the result of Blanchet-Sadri *et al.* stating that the hole must be in the shortest square. In addition, our proof gives a new proof for the original result of Fraenkel and Simpson. Finally, our result implies that the maximum number of squares in a word with one hole is $4n$, regardless of the size of the alphabet. For binary partial words with one hole, we can decrease this bound to $3n$.

2 Preliminaries

We recall some notions and notation mainly from [2]. A word $w = a_1a_2 \cdots a_n$ of length n over an alphabet $\mathcal{A}$ is a mapping $w\colon \{1, 2, \ldots, n\} \to \mathcal{A}$ such that $w(i) = a_i$. The elements of $\mathcal{A}$ are called *letters*. The length of a word w is denoted by $|w|$, and the length of the empty word ε is zero. The set of all finite words including the empty word is denoted by $\mathcal{A}^*$. Let also $\mathcal{A}^+ = \mathcal{A}^* \setminus \{\varepsilon\}$. A word v is a *factor* of a word w (resp. a *prefix*, a *suffix*), if there exist x and y in $\mathcal{A}^*$ such that $w = xvy$ (resp. $w = vy, w = xv$). The prefix (resp. a suffix) of w of length n is denoted by $\mathrm{pref}_n(w)$ (resp. $\mathrm{suf}_n(w)$). The kth *power* of a word $u \neq \varepsilon$ is the word $u^k = \mathrm{pref}_{k\cdot|u|}(u^\omega)$, where u^ω denotes the infinite catenation of the word u with itself and k is a positive rational number such that $k \cdot |u|$ is an integer. If k is an integer, then the power is called an *integer power*. A *primitive* word is a word that is not an integer power of any other word. If $w = uv$, then $u^{-1}w = v$ is the *left quotient* of w by u. If u is not a prefix of w, then $u^{-1}w$ is undefined. Analogously, we define the *right quotient* wu^{-1}.

A partial word u of length n over the alphabet $\mathcal{A}$ is a partial function $u\colon \{1, 2, \ldots, n\} \to \mathcal{A}$. The domain $D(u)$ is the set of positions $i \in \{1, 2, \ldots, n\}$ such that $u(i)$ is defined. The set $H(u) = \{1, 2, \ldots, n\} \setminus D(u)$ is called the set of *holes*. If $H(u)$ is empty, then u is a *full* word. As for full words, we denote by $|u| = n$ the length of a partial word u. Let $\diamond$ be a symbol that does not belong to $\mathcal{A}$. For a partial word u, we define its *companion* to be the full word $u_\diamond$ over the augmented alphabet $\mathcal{A}_\diamond = \mathcal{A} \cup \{\diamond\}$ such that $u_\diamond(i) = u(i)$, if $i \in D(u)$, and $u_\diamond(i) = \diamond$, otherwise. The set $\mathcal{A}_\diamond^*$ corresponds to the set of finite partial words. A partial word u is said to be *contained* in v (denoted by $u \subset v$) if $|u| = |v|$, $D(u) \subseteq D(v)$ and $u(i) = v(i)$ for all $i \in D(u)$. Two partial words u and v are *compatible* (denoted by $u \uparrow v$) if there exists a (partial) word z such that $u \subset z$ and $v \subset z$. In terms of companion words, $u \uparrow v$ if and only if $u_\diamond(i) = v_\diamond(i)$ whenever $u_\diamond(i) \neq \diamond$ and $v_\diamond(i) \neq \diamond$.

For partial words v and w, write $last_w(v) = i$ if $w = u_1v_1u_2$, where $|u_1| = i - 1$, $v \uparrow v_1$ and v is not compatible with a factor in v_1u_2 except the prefix. (If no such v_1 exists, then let $last_w(v)$ be undefined.) If defined, then v_1 is *the last compatible occurrence* of v in w.

A *square* in a partial word is a non-empty factor of the form ww' such that $w \uparrow w'$. If such a square is a full word, then it is called a *full square*. The number of distinct full squares compatible with the factors of a partial word w is denoted by $Sq(w)$:

$$Sq(w) = card\{u^2 \in \mathcal{A}^+ \mid u^2 \uparrow v,\ v \text{ is a factor of } w\}.$$

For each full square u^2 taking part in $Sq(w)$, it suffices to consider the rightmost occurrence of a factor v that is compatible with u^2. Moreover, let

$$s_w(i) = card\{u^2 \in \mathcal{A}^+ \mid last_w(u^2) = i\}.$$

As an example, consider the partial word $w = aba\diamond babaab$ with one hole, $H(w) = \{4\}$. Here $last_w((aba)^2) = 1 = last_w((abaab)^2)$. Also, $(ab)^2$ begins at position 1, but $(ab)^2 \uparrow \diamond bab$, and therefore $last_w((ab)^2) = 4$. Hence $s_w(1) = 2$. Continuing we see that $(s_w(1), s_w(2), \ldots, s_w(|w|)) = (2, 1, 0, 2, 1, 0, 0, 1, 0, 0)$ and therefore we have $Sq(w) = \sum_{i=1}^{|w|} s_w(i) = 7$.

Using the above notation we may state the theorem of Fraenkel and Simpson as follows.

Theorem 1 ([6]). *For any full word $w \in \mathcal{A}^*$, we have $Sq(w) < 2|w|$.*

Since $Sq(w) = \sum_{i=1}^{|w|} s_w(i)$ and no square can start from the last position, i.e., $s_w(|w|) = 0$, the theorem is a direct consequence of the following lemma, which was already proved in a slightly different form in [5]. See also Section 8.1.5 in [12].

Lemma 1. *For any word $w \in \mathcal{A}^*$, we have $s_w(i) \leq 2$ for $i = 1, 2, \ldots, |w|$.*

We finish this section by stating three lemmata which will be needed later in this article. We begin with a characterization for two commuting words. For the proof, see, for example, [11].

Lemma 2. *If $xy = yx$ for full words x and y, then there exists a word z and integers s and t such that $x = z^s$ and $y = z^t$.*

The second lemma, which was proved by Berstel and Boasson in [2], reduces the considerations of partial words to full words as in Lemma 2.

Lemma 3 ([2]). *Let x be a partial word with at most one hole, and let u and v be two (full) words. If $x \subset uv$ and $x \subset vu$, then $uv = vu$.*

Full words u and v are *conjugate* if there exist words x and y such that $u = xy$ and $v = yx$. The third lemma gives a characterization to conjugates using a word equation. For the proof, see, for example, [11].

Lemma 4. *Two words u and v are conjugate if and only if there exists a word z such that $uz = zv$. Moreover, in this case there exist words x and y such that $u = xy$, $v = yx$ and $z = (xy)^n x$ for some integer $n \geq 0$.*

3 Maximum number of last occurrences of squares

Let $card(\mathcal{A}) = k$. In this section we show that, for partial words with one hole, the maximum number of last occurrences of squares starting at the same position is $2k$. For a partial word w and a letter $a \in \mathcal{A}$, we denote by $w(a)$ the full word where the holes are replaced by a. Most certainly $w \subset w(a)$.

Theorem 2. *Let w be a partial word over $\mathcal{A}$ such that w contains only one hole. Then $s_w(i) \leq 2k$, where $k = |\mathcal{A}|$.*

Proof. Suppose that $s_w(i) > 2k$. Each square factor v with $card(H(v)) = 1$, say $v \uparrow u^2$, satisfies $v(a) = u^2$ for a unique letter a filling the single hole. By the pigeon hole principle, there exists a letter $a \in \mathcal{A}$ such that $w(a)$ contains more than two last occurrences of squares starting at the position i. This contradicts with Lemma 1. $\square$

Next we construct recursively a partial word w such that $s_w(1) = 2k$. Let $\mathcal{A} = \{a_1, \ldots, a_k\}$. Let $w_0 = \diamond a_k a_{k-1} \cdots a_1$ and, for $j = 1, 2, \ldots, k$, set

$$\begin{aligned} w_{2j-1} &= w_{2j-2} \cdot w_{2j-2}(a_j), \\ w_{2j} &= w_{2j-1} \cdot (\diamond^{-1} w_{2j-1}) a_j^{-1}, \end{aligned}$$

where the dots emphasize that the substitution is done only in the suffix part. For instance, for $k = 3$, we have $w_0 = \diamond a_3 a_2 a_1$, $w_1 = \diamond a_3 a_2 a_1 a_1 a_3 a_2 a_1$, $w_2 = \diamond a_3 a_2 a_1 a_1 a_3 a_2 a_1 a_3 a_2 a_1 a_1 a_3 a_2$.

One easily shows by induction that $a_k a_{k-1} \cdots a_j$ is a suffix of w_{2j-1}. Also, since w_{2j-1} begins with a hole, the recursive rule with quotients for w_{2j} is well-defined. Notice that

$$w_{2j-1}(a_j) = (w_{2j-2}(a_j))^2 \quad \text{and} \quad w_{2j}(a_j) = (w_{2j-1}(a_j) a_j^{-1})^2.$$

It is clear that $w = w_{2k}$ has a prefix compatible with $w_{2j-1}(a_j) = (w_{2j-2}(a_j))^2$, and a prefix compatible with $w_{2j}(a_j) = (w_{2j-1}(a_j) a_j^{-1})^2$. Therefore we have

Lemma 5. *Let $w = w_{2k}$. Then, for each $j = 1, 2, \ldots, k$, the squares $w_{2j-1}(a_j)$ and $w_{2j}(a_j)$ are prefixes of $w(a_j)$.*

The next lemma shows that the above $2k$ squares do not occur later in w.

Lemma 6. *The full square $w_{2j-1}(a_j)$ is not a factor of w for any $j = 1, 2, \ldots, k$.*

Proof. By the definition, we have

$$w_{2j} = w_{2j-2}(w_{2j-2}(a_j))(\diamond^{-1} w_{2j-2})(w_{2j-2}(a_j) a_j^{-1}).$$

By induction, the set of letters occurring one position before any occurrence of a_k in w_{2j} or in w_{2j-1} is $\{\diamond, a_1, a_2, \ldots, a_j\}$. Hence, $a_j a_k$ does not occur in w_{2j-2}. Since $w_{2j-1}(a_j) = a_j a_k \cdots$, the only possible beginning positions for the factor $w_{2j-1}(a_j)$ in w_{2j} are $l + 1$, $2l$ and $3l$, where $l = |w_{2j-2}|$. However, the factor of length $|w_{2j-1}| = 2l$ at position $l+1$ begins with $w_{2j-2}(a_j) a_k$, which is not a prefix of $w_{2j-1}(a_j)$. Consequently, $w_{2j-2}(a_j)$ does not occur anywhere in w_{2j}, since the positions $2l$ and $3l$ are two close to the end of w_{2j}.

Moreover, $w_{2j-1}(a_j)$ does not occur in $w_{2j+1} = w_{2j} w_{2j}(a_j)$. Namely, the factor of length $2l$ starting at the position $2l$ begins with $w_{2j-2}(a_j)(w_{2j-2}(a_j) a_j^{-1}) a_{j+1}$ and the factor of length $2l$ starting at the position $3l$ begins with $(w_{2j-2}(a_j) a_j^{-1}) a_{j+1}$.

Neither of those are prefixes of $w_{2j-1}(a_j)$. Again, the other possible positions are too close to the end of the word.

By the construction, we conclude inductively that every factor of length $2l$ in $w = w_{2k}$ beginning with $a_j a_k$ has a prefix of the form $w_{2j-2}(a_j)a_k$, $w_{2j-2}(a_j)(w_{2j-2}(a_j)a_j^{-1})b$ or $(w_{2j-2}(a_j)a_j^{-1})b$, where $b \in \{a_{j+1}, a_{j+2}, \ldots a_k\}$. None of these is a prefix of $w_{2j-1}(a_j)$. Hence, $w_{2j-1}(a_j)$ cannot be a factor of w. $\square$

Since $w_{2j-1}(a_j)$ is a prefix of $w_{2j}(a_j)$, we obtain the following corollary.

Corollary 1. *The full square $w_{2j}(a_j)$ is not a factor of w for any $j = 1, 2, \ldots, k$.*

Thus, the previous lemma and the corollary together imply the desired result.

Theorem 3. *For $w = w_{2k}$, we have $s_w(1) = 2k$.*

Note that the above construction for w gives an improvement of the example in [4] containing $k+1$ last compatible occurrences of squares as prefixes. If $k = 2$, our construction gives the binary word of length 38:

$$w = \diamond baababaabbbaababaabbaababaabbbaababaa.$$

The full squares compatible with the prefixes of w are

$$\begin{aligned} w_1(a) &= (aba)^2, \\ w_2(a) &= (abaab)^2, \\ w_3(b) &= (bbaababaab)^2, \\ w_4(b) &= (bbaababaabbbaababaa)^2. \end{aligned}$$

These squares do not occur later in w. Hence, $s_w(1) = 4 = 2k$. Note that here the hole is in the first position. In general, by a result in [4], the hole must be in the shortest last compatible occurrence of a square starting at i whenever $s_w(i) > 2$. As another example, consider the word of length 46:

$$w' = abaab\diamond baabbaabaabbbaabbabaabbbaabbaabaabbbaabb$$

Again $s_w(1) = 4$ and the full squares compatible with the prefixes of w' are $(aba)^2$, $(abaab)^2$, $(abaabbbaabba)^2$ and $(abaabbbaabbaabaabbbaabb)^2$. Now the hole is in the last possible position, namely in the end of the shortest last compatible occurrence of a square starting at position one.

4 Distinct squares in a partial word with one hole

In this section our goal is to estimate how many distinct squares can occur in a partial word with one hole. We start by proving the following technical lemma. In the sequel, we denote

$$w[i, j] = w(i)w(i+1)\cdots w(j)$$

for a word w and integers i and j with $i < j$. The integer part of a real number x is denoted by $\lfloor x \rfloor$.

Lemma 7. *Let vv' be a prefix of ww', where $v \uparrow v'$, $w \uparrow w'$ such that $|w| < 2|v|$, say $l = |w| - |v| < |v|$. Assume that ww' contains at most one hole and denote by V the full word compatible with both v and v'. Then there are words Z and $\widehat{Z}$ of length l such that*

$$V = \begin{cases} Z^{m+1}\widehat{Z}^n\widehat{Z}_1 & \text{if } v(h) = \diamond \text{ with } 1 \leq h \leq l\lfloor |v|/l \rfloor, \\ Z^m\widehat{Z}^n\widehat{Z}_1 & \text{if } v'(h) = \diamond \text{ with } l+1 \leq h \leq |v|, \\ \widehat{Z}^n\widehat{Z}_1 & \text{otherwise}, \end{cases} \tag{1}$$

where $m = \lfloor (h-1)/l \rfloor$, n is a non-negative integer, $\widehat{Z}_1$ is a prefix of $\widehat{Z}$, and there exists a partial word z containing at most one hole and satisfying $z \subset Z$ and $z \subset \widehat{Z}$.

Proof. Let us first consider the case where $v(h) = \diamond$ and $1 \leq h \leq l\lfloor |v|/l \rfloor$. Consider a non-negative integer k such that $(k+1)l \leq |v|$. Since $v \uparrow v'$, we have

$$v[kl+1,(k+1)l] \subset v'[kl+1,(k+1)l]. \tag{2}$$

Moreover, since $l < |v|$, the word w' begins inside v' at the position $l+1$ and, therefore, $w'[kl+1,(k+1)l] = v'[(k+1)l+1,(k+2)l]$, if $(k+2)l \leq |v'| = |v|$. Since $w \uparrow w'$, we have $v[kl+1,(k+1)l] = w[kl+1,(k+1)l] \subset w'[kl+1,(k+1)l]$. Hence, combining these facts, we obtain

$$v[kl+1,(k+1)l] \subset v'[(k+1)l+1,(k+2)l]. \tag{3}$$

If $(k+1)l < |v| < (k+2)l$, it is clear that (3) holds for prefixes of length $|v| - (k+1)l$ of the considered words. Note that the relation $\subset$ occurring in both equations can be replaced by the identity relation whenever $v[kl+1,(k+1)l]$ is a full word. Hence, applying (2) and (3) for different values of k, we conclude that $V = v' = Z^{m+1}\widehat{Z}^n\widehat{Z}_1$, where $m = \lfloor (h-1)/l \rfloor$, $Z = v'[ml+1,(m+1)l]$, $z = v[ml+1,(m+1)l]$ and, we may choose

$$\widehat{Z} = \begin{cases} v'[(m+1)l+1,(m+2)l] & \text{if } (m+2)l \leq |v'|, \\ (v'[(m+1)l+1,|v'|])(v'[|v'|-l+1,(m+1)l]) & \text{otherwise.} \end{cases}$$

In the case where $v'(h) = \diamond$ and $l \leq h \leq |v'|$, we notice that instead of (2) and (3) the following equations hold:

$$v'[(k+1)l+1,(k+2)l] \subset v[(k+1)l+1,(k+2)l] \tag{4}$$

and

$$v'[(k+1)l+1,(k+2)l] \subset v[kl+1,(k+1)l]. \tag{5}$$

Similarly to the first case, we conclude using (4) and (5) that $V = v = Z^m \widehat{Z}^n \widehat{Z}_1$, where $m = \lfloor (h-1)/l \rfloor \geq 1$, $Z = v[(m-1)l+1, ml]$,

$$z = \begin{cases} v'[ml+1, (m+1)l] & \text{if } (m+1)l \leq |v'|, \\ (v'[ml+1, |v'|])(v'[|v'|-l+1, ml]) & \text{otherwise.} \end{cases}$$

and

$$\widehat{Z} = \begin{cases} v[ml+1, (m+1)l] & \text{if } (m+1)l \leq |v|, \\ (v[ml+1, |v|])(v[|v|-l+1, ml]) & \text{otherwise.} \end{cases}$$

If the hole occurs in $v[l\lfloor |v|/l \rfloor + 1, |v|]$, then set $\widehat{Z} = v'[1, l]$ and use (2) and (3) with identity relation instead of $\subset$ to obtain $V = v' = \widehat{Z}^n \widehat{Z}_1$. If the hole occurs in $v'[1, l]$, then set $\widehat{Z} = v[1, l]$ and apply (4) and (5). If the word vv' is full, the result is obvious. □

Our next result concerns the lengths of squares starting at the same position. This theorem has a crucial role in the sequel. Namely, if $s_w(i) > 2$ for some position i in w, then the theorem says that the suffix of w starting at i must be quite long. Hence, the maximum value of $s_w(i)$ is dependent on how far the position i is from the end of the word w. This restricts the total number of distinct squares compatible with the factors of a partial word.

Theorem 4. *If three distinct full squares have their last compatible occurrences in a partial word with one hole starting at the same position, then the longest square is at least twice as long as the shortest square.*

Proof. Consider a partial word with one hole. Assume that three partial words uu', vv' and ww', where $u \uparrow u'$, $v \uparrow v'$, $w \uparrow w'$, and $|u| = p < |v| = q < |w| = r$, start at the same position in the word. Denote by U^2 (resp. V^2, W^2) the full word that contains uu' (resp. vv', ww'). Assume also that U^2 (resp. V^2, W^2) is not compatible with any factor occurring later in the word. Moreover, assume that $r < 2p$, i.e., $r - p < p$. Denote the position of the hole in ww' by h. We divide the proof into three cases:

$$\text{A. } h \notin [1, r-p], \quad \text{B. } h \in [1, q-p], \quad \text{C. } h \in [q-p+1, r-p].$$

Case A. Assume that $h \notin [1, r-p]$. Since $r - p < p < q$, there exist words $U[1, r-p]$, $V[1, r-p]$ and $W[1, r-p]$ and, by the assumption, these words are equal to $u[1, r-p]$. Let $X = U[1, r-p] = X_1X_2$, where $X_1 = U[1, q-p]$. Since $v'[1, r-p] \subset V[1, r-p] = U[1, r-p]$, we have $v'[1, r-p] \subset X_1X_2$. Similarly, we also have $u'[1, r-p] \subset X_1X_2$ and $w'[1, r-p] \subset X_1X_2$. Hence, $v'[1, r-p]$ is contained both in X_1X_2 and in X_2X_1; see Figure 1. By Lemma 3, we have $X_1X_2 = X_2X_1$ and, by Lemma 2, there exists a full word Y such that both X_1 and X_2 are integer powers of Y.

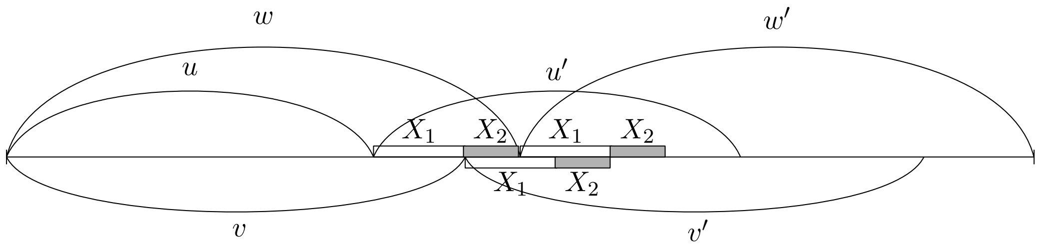

Figure 1: Illustrations of three partial squares uu', vv' and ww' starting at the same position and satisfying $|u| = |u'| < |v| = |v'| < |w| = |w'| < |u^2|$.

Since w' starts inside v' at the position $|X_2| + 1$, we may use Lemma 7 and we notice that in all cases of (1) the full word V can be written in the form $(X_2)^m(\widehat{X}_2)^n\widehat{X}_2'$, where m and n are suitably chosen integers, $\widehat{X}_2 = \widehat{X}_2'\widehat{X}_2''$ and there exists a partial word with one hole contained in both X_2 and $\widehat{X}_2$. Hence, there is at most one position where X_2 and $\widehat{X}_2$ may differ. Moreover, since X_2 is an integer power of Y, it follows that $X_2 = Y^k$ and $\widehat{X}_2 = Y^i\widehat{Y}Y^j$, where $i+j+1 = k$ and there exists a partial word y with one hole compatible with both Y and $\widehat{Y}$.

Now consider the word $z = \mathrm{suf}_{|Y|}(v)$. Let us denote $Y = Y_1Y_2$, $\widehat{Y} = \widehat{Y}_1\widehat{Y}_2$ and $y = y_1y_2$, where $|Y_1| = |\widehat{Y}_1| = |y_1| = q - \lfloor q/|Y| \rfloor |Y|$. Since a partial word y with only one hole satisfies $y \subset Y_1Y_2$ and $y \subset \widehat{Y}_1\widehat{Y}_2$, we conclude that either $\widehat{Y}_1 = Y_1$ and the word y_2 with one hole satisfies $y_2 \subset Y_2$ and $y_2 \subset \widehat{Y}_2$ or $\widehat{Y}_2 = Y_2$ and the word y_1 with one hole satisfies $y_1 \subset Y_1$ and $y_1 \subset \widehat{Y}_1$. Hence, by the form of V, we have three possibilities: (a) $z \subset \widehat{Y}_2Y_1$, (b) $z \subset Y_2\widehat{Y}_1$, and (c) $z \subset Y_2Y_1$. On the other hand, since $u'[1, |X_1|] \subset X_1$ and X_1 is an integer power of Y, we have $z \subset Y = Y_1Y_2$.

Suppose first that z is a full word and consider the case (a); see Figure 2. Now the full word z is equal to $\widehat{Y}_2Y_1 = Y_1Y_2$ and we may use Lemma 4 to conclude that $\widehat{Y}_2 = Z_1Z_2$, $Y_1 = (Z_1Z_2)^rZ_1$ for some integer r, and $Y_2 = Z_2Z_1$. However, we know that there exists the word y_2 which is contained in both $\widehat{Y}_2 = Z_1Z_2$ and $Y_2 = Z_2Z_1$. By Lemma 3, this means that $Z_1Z_2 = Z_2Z_1$ and, by Lemma 2, there exists a word α such that both Z_1 and Z_2 are integer powers of α. Moreover, it follows that Y_1, $Y_2 = \widehat{Y}_2$, and consequently, $Y = \widehat{Y}$ are integer powers of α. Hence, also the words V, X_1 and U are integer powers of this word. This means that the word $vv'[1 + |\alpha|, 2p + |\alpha|] \subset VV[1, 2p] = U^2$. Thus, uu' is not the last compatible occurrence of U^2, which is a contradiction.

If z is a full word, case (b) is symmetric to case (a). In case (c), we immediately have $Y_1Y_2 = Y_2Y_1$ and, by Lemma 2, both Y_1 and Y_2 are powers of the same word. This gives a contradiction the same way as above.

Suppose next that there is a hole in z, i.e., $v(h) = \diamond$ for some position h. Denote $l = \lfloor |v|/|X_2| \rfloor$ and $z = z_2z_1$, where $|z_1| = |Y_1|$. Recall that $V = (X_2)^m(\widehat{X}_2)^n\widehat{X}_2'$, where $X_2 = Y^k$ and $\widehat{X_2} = Y^i\widehat{Y}Y^j$.

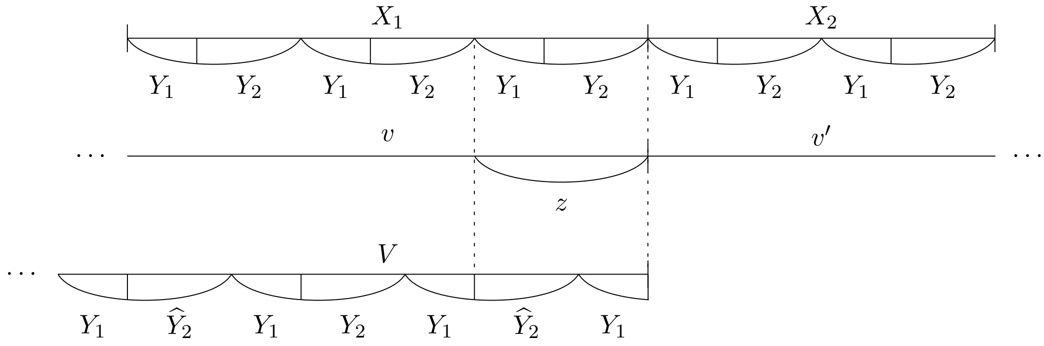

Figure 2: Illustration of case (a), where $X_1 = (Y_1Y_2)^3$, $X_2 = (Y_1Y_2)^2$, $\widehat{X}_2 = (Y_1\widehat{Y}_2)(Y_1Y_2)$ and $\widehat{X}'_2 = (Y_1\widehat{Y}_2)Y_1$.

Assume first that the hole occurs in a non-empty partial word z_1. Since $|X_2|$ and $|v[1, q - |Y_1|]|$ are multiples of $|Y|$, this means that $h > l|X_2|$ and, by Lemma 7, we obtain $V = \widehat{X}_2^{n'}\widehat{X}'_2$, where $n' = m + n$. Thus, this means that $X_2 = \widehat{X}_2$, $Y = \widehat{Y}$ and, consequently, $z \subset Y_2Y_1$. Since we have shown above that also $z \subset Y_1Y_2$, we may use Lemma 3 and Lemma 2 to conclude that Y_1, Y_2 are powers of some word α. Consequently, the words Y, X_2 and X_1 are non-empty powers of α. Since the prefix $\widehat{X}'_2$ of $X_2 = \widehat{X}_2$ must be of the form $(Y_1Y_2)^iY_1$, we conclude that also V is a power of α. Thus, $U = VX_1^{-1}$ is a power of α and there is a word compatible with U^2 beginning at position $\alpha + 1$ in vv'. This is a contradiction.

Finally, assume that the hole occurs in z_2. If $h > l|X_2|$, we get a contradiction as above. However, it is now possible that $h \in [(l-1)|X_2| + 1, l|X_2|]$. By Lemma 7, this means that $V = X_2^{m'}\widehat{X}'_2$, where $m' = m + n$. Moreover, z_2 must be a suffix of $v[(l-1)|X_2| + 1, l|X_2|]$ and $\widehat{X}'_2$ is the full word z_1. By the form of V, we conclude that

$$v[(l-1)|X_2| + 1, l|X_2|] \subset X_2 = Y^k. \tag{6}$$

Hence, it follows that $z_2 \subset Y_2$. On the other hand, the proof of Lemma 7 gives

$$v[(l-1)|X_2| + 1, l|X_2|] \subset \widehat{X}_2 = \widehat{X}'_2\widehat{X}''_2. \tag{7}$$

Let us denote $X_2 = X'_2X''_2$, where $|X'_2| = |\widehat{X}'_2|$. Recall that $\widehat{X}'_2 = z_1$ and $|z_1| = |Y_1|$. Since the length of $v[(l-1)|X_2| + 1, l|X_2|]$ is a multiple of $|Y|$, the hole occurring in the suffix z_2 cannot occur in $\text{pref}_{|Y_1|}(v[(l-1)|X_2| + 1, l|X_2|])$. Therefore, it follows by (6) and (7) that $z_1 = \widehat{X}'_2 = X'_2 = Y_1$. Hence, we have shown that $z = z_2z_1 \subset Y_2Y_1$. Since also $z \subset Y_1Y_2$, we use Lemma 3 and Lemma 2 to conclude that Y_1, Y_2 and, consequently, Y, X_2 and X_1 are powers of some word α. Thus, the prefix $V[1, q - |Y_1|] = X_2^{m'}$ must be a power of α. Since $q - |Y_1| \geq p$, it follows that $U = V[1, p] = \alpha^t$ for some positive integer t. Recall that $v[p+1,q] \subset X_1$, where $X_1 = \alpha^s$ for some positive integer s. Hence, the partial word $v[1,p]v[p+1,q]v'[1,p]$ is contained in $UX_1U = \alpha^{t+s+t}$. Thus, there is a factor of vv' compatible with $U^2 = \alpha^{2t}$ starting at position $\alpha+1$. Once again we end up in a contradiction.

Case B. Assume that the hole occurs in $u[1,q-p]$. Hence, we may denote $U[1,r-p] = X_1X_2$, $V[1,r-p] = \widetilde{X}_1X_2$ and $W[1,r-p] = \widehat{X}_1X_2$, where $u[1,q-p]$ is contained in X_1, $\widetilde{X}_1$ and $\widehat{X}_1$. Since $u'[q+p+1,r-p] = X_2$ and $w'[1,q-p] = \widehat{X}_1$, it follows that $v'[1,r-p] = \widetilde{X}_1X_2 = X_2\widehat{X}_1$. By Lemma 4, there exist Z_1 and Z_2 such that $\widetilde{X}_1 = Z_1Z_2$, $\widehat{X}_1 = Z_2Z_1$ and $X_2 = (Z_1Z_2)^rZ_1$ for some integer r. Since $u[1,q-p] \subset \widetilde{X}_1$ and $u[1,q-p] \subset \widehat{X}_1$, it follows that $Z_1Z_2 = Z_2Z_1$ by Lemma 3. Hence, by Lemma 2, there exists a word Y such that $\widetilde{X}_1 = \widehat{X}_1 = Y^k$ and $X_2 = Y^l$. Since there is only one hole in $u[1,q-p]$ and $u[1,q-p]$ is contained in both X_1 and $\widetilde{X}_1$, we may write $X_1 = Y^i\widehat{Y}Y^j$, where $i+j+1 = k$ and there is a word y with one hole compatible with both Y and $\widehat{Y}$.

Using the proof of Lemma 7, we conclude that $V = X_2^{m+1}\widehat{X}_2^n\widehat{X}_2'$, where X_2 and $\widehat{X}_2$ are compatible, m and n are integers, and the hole occurs in $v[m|X_2|+1,(m+1)|X_2|]$. Since the position of the hole in v is at most $q-p = |X_1|$, it follows that $m|X_2| < |X_1|$. Since $|X_1| = |\widehat{X}_1| = k|Y|$ and $|X_2| = |\widehat{X}_2| = l|Y|$, we have $|X_2^{m+1}\widehat{X}_2| = (m+2)l|Y| < (2l+k)|Y| = |X_2\widehat{X}_1X_2|$. Hence, the word $X_2^{m+1}\widehat{X}_2$ is a prefix of $v'[1,2r-q-p] = u'[q-p+1,r-p]w'[1,r-p] = X_2\widehat{X}_1X_2 = Y^{2l+k}$. Since $|X_2| = |\widehat{X}_2| = l|Y|$, it follows that $X_2 = \widehat{X}_2 = Y^l$ and $V = Y^{n'}Y_1$, where n' is an integer and $Y = Y_1Y_2$.

As in the previous case, consider the word $z = \mathrm{suf}_{|Y|}(v)$ and denote $\widehat{Y} = \widehat{Y}_1\widehat{Y}_2$ and $y = y_1y_2$, where $|Y_1| = |\widehat{Y}_1| = |y_1|$. Recall that y is a word with one hole contained in both Y and $\widehat{Y}$. Hence, the hole is either in y_1 or y_2, and consequently, we have either $\widehat{Y} = \widehat{Y}_1Y_2$ or $\widehat{Y} = Y_1\widehat{Y}_2$. Since $u'[1,q-p] = X_1 = Y^i\widehat{Y}Y^j$ is a suffix of V, there are three possibilities: (a) $z = \widehat{Y}_1Y_2$ (b) $z = Y_1\widehat{Y}_2$ (c) $z = Y_1Y_2$. On the other hand, the structure of V implies that $z = Y_2Y_1$ and, as in Case A, all the subcases (a)–(c) lead to a contradiction.

Case C. Assume that the hole occurs in $u[q-p+1,r-p]$. Now we have $U[1,r-p] = X_1X_2$, $V[1,r-p] = X_1\widetilde{X}_2$ and $W[1,r-p] = X_1\widehat{X}_2$, where $u[q-p+1,r-p]$ is contained in X_2, $\widetilde{X}_2$ and $\widehat{X}_2$. Since $u'[q+p+1,r-p] = X_2$ and $w'[1,q-p] = X_1$, it follows that $v'[1,r-p] = X_1\widetilde{X}_2 = X_2X_1$. By Lemma 4, there exists Z_1 and Z_2 such that $X_2 = Z_1Z_2$, $\widetilde{X}_2 = Z_2Z_1$ and $X_1 = (Z_1Z_2)^rZ_1$ for some integer r. Since $u[q-p+1,r-p]$ is contained in $X_2 = Z_1Z_2$ and $\widetilde{X}_2 = Z_2Z_1$, we conclude by Lemma 3 and Lemma 2 that X_2 and X_1 are integer powers of some full word Y. We get a contradiction exactly the same way as in Case A.

□

As a corollary, we get the following result.

Corollary 2 ([4]). *If three distinct squares have their last compatible occurrences in a partial word with one hole starting at the same position, then the hole must be in the shortest square.*

Proof. Let z be a partial word with one hole. Assume that uu', vv' and ww', where $u \uparrow u'$, $v \uparrow v'$ and $w \uparrow w'$, begin at the same position in z. Let these partial words be the last compatible occurrences of three distinct full squares in z. Assume also that uu' is a full word, i.e., $u = u'$. This implies that $|w| < |u^2|$ as otherwise a word compatible with u^2 would appear later in the word. By Theorem 4, this is impossible. Hence, the hole must be in the shortest square. □

Moreover, our proof for Theorem 4 gives a new proof for the original theorem of Fraenkel and Simpson (Theorem 1). Note that the proof can be considerably shortened and simplified if the words do not contain any holes.

Next we use Theorem 4 to show that the number of distinct squares in a partial word with one hole does not depend on the size of the alphabet. This may be surprising since the maximum of $s_w(i)$ is dependent on the alphabet size as was shown in the previous section.

Theorem 5. *For any partial word w with one hole, we have $Sq(w) < 4|w|$.*

Proof. Suppose that $w(j) = \diamond$ and denote $n = |w|$. If $s_w(i) = 3$, then $i < j$ and the last compatible occurrence of the shortest square must contain a hole by Corollary 2. Hence, the length of the shortest square is at least $j - i + 1$. By Theorem 4, the suffix of w beginning after the hole must be at least as long as the shortest square. Thus, we must have $n - j > j - 1 + i$. If $s_w(i) = 4$, Theorem 4 does not give much information as already the second largest square is twice as long as the second shortest. However, if $s_w(i) = 5$, we may consider only the three largest squares, where the shortest one is of size $2(j - 1 + i)$. Hence, the longest square must be twice as long as the shortest and, therefore, $n - j > 3(j - 1 + i)$. By induction, we conclude that $n-j > (2^{k-1}-1)(j-i+1)$ whenever $s_w(i) = 2k$. In other words, we have an estimate $s_w(i) \leq 2k$, where

$$k = 1 + \log_2 \left(\frac{n - i + 1}{j - i + 1} \right). \tag{8}$$

Note that here we assume that the size of the alphabet is large enough. Hence, (8) gives us an upper bound on the number of distinct squares in w.

$$Sq(w) \leq 2 \sum_{i=1}^{j} \left(1 + \log_2 \left(\frac{n - i + 1}{j - i + 1} \right) \right) + 2(n - j - 1).$$

Here the last term $2(n - j - 1)$ corresponds to the positions after the hole. There, we have $s_w(i) \leq 2$ for $i = j, j + 1, \ldots, n - 1$ by Theorem 1, and the last position cannot contain any squares. Using the natural logarithm $\ln$, we may write

$$Sq(w) \leq \frac{2}{\ln 2}\left(\sum_{i=1}^{j} \ln(n-i+1) - \sum_{i=1}^{j} \ln(j-i+1)\right) + 2n - 2.$$

Since $\ln(n-i+1)$ and $\ln(j-i+1)$ are strictly decreasing in i, we may estimate that $Sq(w) \leq f(j)$, where

$$f(j) = 2\left(\ln n + \int_1^j \ln(n-x+1)\,\mathrm{d}x - \int_0^{j-1} \ln(j-x)\,\mathrm{d}x\right) + 2n - 2.$$

By integrating, we obtain

$$f(j) = 2(\ln n - (n-j+1)\ln(n-j+1) + n\ln n - j\ln j) + 2n - 2.$$

The maximum value of $f(j)$ in the interval $[1, n-2]$ is obtained at the critical point $j = (n+1)/2$, where

$$f(j) = 4n + \frac{1+n}{\ln 2}\ln\left(\frac{n}{1+n}\right) \leq 4n - \frac{2}{\ln 2}.$$

Note that if $j > n-2$, then $Sq(w) = 2n$. Hence, we have proved that $Sq(w) < 4n$ regardless of the size of the alphabet. □

Of course, we get better estimates if the size of the alphabet is restricted. As a final theorem, let us consider binary words.

Theorem 6. *For any binary partial word w containing one hole, we have $Sq(w) \leq 3n$.*

Proof. Suppose that $w(j) = \diamond$. Since w is a binary partial word, Theorem 2 implies that $s_w(i) \leq 4$ for every position i. On the other hand, Theorem 4 restricts the number of position, where $s_w(i) > 2$. If $j \geq n/2$, then these positions are in the interval $[2j-n+1, j]$. Hence, $(n-j)$ positions may have $s_w(j) = 4$. This gives us

$$Sq(w) \leq 4(n-j) + 2j = 4n - 2j \leq 3n, \tag{9}$$

since $j \in [n/2, n]$. Similarly, if $j < n/2$, then $s_w(i) = 4$ is possible only for position i in the interval $[1, j]$. Thus, for $j \in [1, n/2)$, we obtain

$$Sq(w) \leq 4j + 2(n-j) = 2n + 2j < 3n. \tag{10}$$

Hence, by inequalities (9) and (10), the claim follows. □

References

[1] J.-P. Allouche, J. Shallit, The ubiquitous Prouhet-Thue-Morse sequence, Springer Ser. Discrete Math. Theor. Comput. Sci., Springer, London, 1999, 1–16.

[2] J. Berstel, L. Boasson, Partial words and a theorem of Fine and Wilf, Theoret. Comput. Sci. 218 (1999) 135–141.

[3] F. Blanchet-Sadri, Algorithmic Combinatorics on Partial Words, Chapman & Hall/CRC Press, Boca Raton, FL, 2007.

[4] F. Blanchet-Sadri, R. Mercaş, G. Scott, Counting distinct squares in partial words. In E. Csuhaj-Varju and Z. Esik (Eds.), AFL 2008, 12th International Conference on Automata and Formal Languages, May 27-30, 2008, Balatonfured, Hungary, Proceedings, pp 122-133.

[5] M. Crochemore, W. Rytter, Squares, cubes, and time-space efficient string searching, Algorithmica 13 (1995), 405–425.

[6] A. S. Fraenkel, J. Simpson, How many squares can a string contain?, J. Combin. Theory Ser. A, 82(1) (1998) 112–120.

[7] V. Halava, T. Harju, T. Kärki, P. Séébold, Overlap-freeness in infinite partial words, to appear in Theoret. Comput. Sci.

[8] V. Halava, T. Harju, T. Kärki, Square-free partial words, Inform. Process. Lett. 108 (2008) 290–292.

[9] L. Ilie, A simple proof that a word of length n has at most $2n$ distinct squares, J. Combin. Theory Ser. A 112 (2005) 163–164.

[10] L. Ilie, A note on the number of squares in a word, Theoret. Comput. Sci. 380 (2007) 373–376.

[11] M. Lothaire, Combinatorics on Words, Encyclopedia of Mathematics 17, Addison-Wesley, 1983.

[12] M. Lothaire, Algebraic combinatorics on words, Encyclopedia of Mathematics and its Applications 90, Cambridge University Press, Cambridge, 2002.

[13] F. Manea, R. Mercaş, Freeness of partial words, Theoret. Comput. Sci. 389(1-2) (2007) 265–277.

[14] A. Thue, Über unendliche Zeichenreihen, Norske Vid. Skrifter I Mat.-Nat. Kl., Christiania 7 (1906) 1–22.

[15] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, Norske Vid. Skrifter I Mat.-Nat. Kl., Christiania 1 (1912) (1–67).

Turku Centre *for* Computer Science

Lemminkäisenkatu 14 A, 20520 Turku, Finland | www.tucs.fi

University of Turku
- Department of Information Technology
- Department of Mathematics

Åbo Akademi University
- Department of Computer Science
- Institute for Advanced Management Systems Research

Turku School of Economics and Business Administration
- Institute of Information Systems Sciences

ISBN 978-952-12-2102-6
ISSN 1239-1891