This item was submitted to [Loughborough's Research Repository](#) by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

Clusters of repetition roots: single chains

PLEASE CITE THE PUBLISHED VERSION

https://doi.org/10.1007/978-3-030-67731-2_29

PUBLISHER

Springer

VERSION

AM (Accepted Manuscript)

PUBLISHER STATEMENT

The final authenticated version is available online at https://doi.org/10.1007/978-3-030-67731-2_29.

LICENCE

CC BY-NC-ND 4.0

REPOSITORY RECORD

Fazekas, Szilárd Zsolt, and Robert Mercas. 2021. "Clusters of Repetition Roots: Single Chains".
Loughborough University. <https://hdl.handle.net/2134/13990997.v1>.

Clusters of repetition roots: single chains

Szilárd Zsolt Fazekas^{1*} and Robert Mercas²

¹ Akita University, Graduate School of Engineering Science, Japan
`szilard.fazekas@ie.akita-u.ac.jp`

² Loughborough University, Department of Computer Science, UK
`R.G.Mercas@lboro.ac.uk`

Abstract. This work proposes a new approach towards solving an over 20 years old conjecture regarding the maximum number of distinct squares that a word can contain. To this end we look at clusters of repetition roots, that is, the set of positions where the root u of a repetition u^ℓ occurs. We lay the foundation of this theory by proving basic properties of these clusters and establishing upper bounds on the number of distinct squares when their roots form a chain with respect to the prefix order.

1 Introduction

Repetitions (periodicities) in words are well-studied, due primarily to their importance in word combinatorics [19] as well as in various applications such as string matching algorithms [6], molecular biology [11], or text compression [21]. The most basic repetitive structure is xx , where x is a non-empty string. Such a string is also called, due to its form $xx = x^2$, a *square*.

A string is said to be square-free or repetition-free if it contains no squares. It was shown by Thue [22, 23] that there exist square-free, respectively, cube-free, strings of infinite length over a ternary, respectively, binary, alphabet. On the other hand, it has been shown that the minimal number of distinct squares that any sufficiently long binary string must contain is three [9].

A string of length n can have $\Theta(n^2)$ squares, by the trivial example of a unary word, and it is known that the maximum number of squares xx , where x itself is not a repetition is $\Theta(n \log n)$ [6]. Repetition counting has also been investigated in other settings: when the length of the root (x for a repetition x^ℓ) has length as small or large as possible (e.g., [7, 9, 20]), for partial words, where words contain extra joker symbols that match every letter of the alphabet (e.g., [2, 3], as well as for abelian and other types of repetitions where the consecutive factors are not identical copies but equivalent in a looser sense (e.g. [17]).

Some, quite old and well studied, problems regarding this topic refer to the maximal number of distinct repetitions that a word can have, as well as to the maximum number of maximal repetitions (runs) that a string can contain.

* This Work Was Supported By JSPS KAKENHI Grant Number JP19K11815.

All repetitions whose roots share a common prefix are in one group. Then, the aim is to show that for every element in a group there are at least a certain number of positions which are not part of the positions of another element's.

§§ **Our Contribution.** In this work we pose a conjecture which would imply Conjecture 1 to be true, and prove it in the special case when the roots of repetitions considered form a chain of square roots (totally ordered set) with respect to the prefix ordering.

It is worth noting that while most of the earlier results concerning the bounds on the maximum number of distinct squares were obtained using some version of the so called Three Squares Lemma [6], the bounds concerning runs, as well as those concerning bounds on the repetitions with integer exponents higher than 2, made use of Lyndon trees and Lyndon words, and these approaches were never connected. In this work, we show the first framework which may allow a unified presentation of the bounds for distinct repetitions and runs by representing repetitions as all positions where their roots (appropriately defined for each particular type of repetition) occur.

§§ **Preliminaries.** A *word* is a concatenation of letters from a *finite alphabet* Σ of size $|\Sigma|$. The *empty word* ε is the word of length 0. For a factorization $w = xyz$, we call x a *prefix* (denoted by $x \leq_p w$, or $x <_p w$ if $x \neq w$) and z a *suffix* of w , while each of x, y, z are called *factors* of w . A factor is *proper* if it is non-empty and not equal to w . If x is both a prefix and suffix of w , then x is a *border* of w . We call p a *period* of w if the letters p positions apart in w are the same. The *minimal period* is given by the smallest such p . By $|w|_x$ we denote the number of times x occurs as a factor of w (including overlaps).

A *repetition* represents consecutive concatenations of the same word. An ℓ -*power* (ℓ -*repetition*) represents ℓ such repetitions of the same factor. If a word is not a repetition, then it is called *primitive*. Moreover, if $w = u^\ell$ is an ℓ -repetition we say that u is a *root* of w , and call u the *primitive root* of w when u is primitive. Finally, t^ω denotes the infinite word formed by consecutive repetitions of t .

Although unnecessary for the proofs, in order to simplify the illustrations, for all words w we consider their *suffix array* structures, which are permutations of $\{1, \dots, |w|\}$ such that the i th entry is the starting position of the i th smallest suffix of w in the lexicographical order.

2 Clusters of repetition roots

In this section we introduce the clusters of repetition roots and prove some fundamental properties for these clusters in relation with clusters of other repetitions. Let us fix a word w and let its suffix array be S .

For each factor u^ℓ of w , we denote by $\mathbf{clust}_w(u)$ the *cluster* in S that contains the starting position of all suffixes having u as a prefix. Fig. 2 illustrates how these clusters could be perceived (for $\ell = 2$), arranged one on top of the other.

Fig. 3. Single (left) and multiple (right) chains of clusters

Observation 1 *The set of suffixes of a word sharing a common prefix are contiguous in the suffix array, forming a cluster. If an ℓ -repetition u^ℓ is a factor of a word, then the suffix array of the word contains a cluster of size at least ℓ of suffixes having the root u of the ℓ -repetition as a prefix.*

As every word, and therefore every suffix having v as prefix, also has u as prefix when $u \leq_p v$, the next observation is straightforward.

Observation 2 *For any two factors u and v of any word w , we have $u \leq_p v \Leftrightarrow \text{clust}_w(v) \subseteq \text{clust}_w(u) \Leftrightarrow \text{clust}_w(u) \cap \text{clust}_w(v) \neq \emptyset$ and $|u| \leq |v|$.*

We pose the following conjecture, which, if true, would give a general upper bound for integer exponent distinct repetitions:

Conjecture 2. For any word w , any positive integer ℓ , and any set of words $S = \{u_1, u_2, \dots, u_n\}$ such that, for all $i \in \{1, \dots, n\}$, u_i^ℓ is a factor of w and $u_1 \leq_p u_i$, we have $|S| < \frac{1}{\ell-1} |w|_{u_1}$.

In other words, for a number n of ℓ -repetitions $u_1^\ell, \dots, u_n^\ell$ with a common prefix x , we conjecture $n < \frac{1}{\ell-1} \cdot |\text{clust}(x)|$. In this paper we approach the problem by analysing the case where $\ell = 2$ and $u_1 \leq_p \dots \leq_p u_n$, that is, S is a set of roots of distinct squares, totally ordered by the prefix relation. We call such a collection of square roots a (*prefix*) *chain*. In Section 3 we prove a special case of Conjecture 2:

Problem 1. For a word w and a prefix chain $S = \{u_1, u_2, \dots, u_n\}$ with $u_i \leq_p u_{i+1}$ for all $i \in \{1, \dots, n-1\}$, we have $|S| < \frac{1}{\ell-1} |w|_{u_1}$.

Compared to any of the results in [10, 15, 16, 8], the bound in Problem 1 is different because it is in a sense optimal, as we will argue at the end of Section 3. It is also important to note that while all the bounds on distinct repetition would be a direct corollary of Conjecture 2, the converse does not hold.

The next lemmas explore the situation of two equal clusters. This is a crucial issue, since when the clusters are all different, the bound easily follows: one can assign to each root a position which is in its cluster but not in the clusters

of longer words. Most of the following results can be proved in general for ℓ -repetitions, but to simplify the exposition, here we usually restrict ourselves to $\ell = 2$. First recall the following well-known result about primitivity of words.

Lemma 1. [19] *A word w is primitive if and only if it occurs only twice in ww .*

In the following lemma we look at the relative positions of the rightmost occurrences of two squares whose roots have the same cluster.

Lemma 2. *For two squares u^2, v^2 with $u <_p v$, if $\mathbf{clust}_w(u) = \mathbf{clust}_w(v)$, then the rightmost occurrences of u^2 and v^2 in w cannot start on the same position.*

Proof. Let the rightmost occurrences of u^2 and v^2 start at the same position i . This means that $|u^2| > |v|$, as otherwise u^2 would occur later, at $i + |v|$. We have occurrences of u at $i, i + |u|$ and $i + |v|$. Since $\mathbf{clust}_w(u) = \mathbf{clust}_w(v)$, we also have v at $i + |u|$ and by Lemma 1, we get that v is non-primitive, and by the theorem of Fine and Wilf, the primitive roots of u and v are the same, say t . Since u is shorter than v , this gives an occurrence of u^2 at $i + |t|$, contradicting the assumption that u^2 does not occur after position i . $\square$

Next, we show that if the clusters of two words are equal and one of the words is non-primitive, then the other word has to be of a very specific form.

Lemma 3. *Consider the factors $u^2 \neq v^2$ of some word w , such that $\mathbf{clust}_w(u) = \mathbf{clust}_w(v)$ and $u <_p v$. Then the following hold:*

1. *if $u = t^m$ for some $m > 1$ and primitive t , then $v = t^m t'$ for $\varepsilon \neq t' \leq_p t$;*
2. *if $v = t^m$ for some $m > 1$ and primitive t , then $u = t^{m-1} t'$ for $\varepsilon \neq t' \leq_p t$.*

Proof. If two words have equal clusters, their rightmost occurrences start at the same position. Let us denote this position by i for u and v .

If $u = t^m$ for some primitive t and $m > 1$, and $|v| \geq 2|u|$, then we get a contradiction, because u occurs at position $i + |u|$. So, $|v| < 2|u|$, which means that $v = t^n t'$, for some $n \geq m$ and $t' \leq_p t$. If $n > m$, then we have an occurrence of u at $i + |t|$. So, $n = m$ and because $u \neq v$, we get that t' is not empty.

If $v = t^m$ for some primitive t and $m > 1$, since $u \leq_p v$ we get that $u = t^n t'$ for $n < m$ and $t' \leq_p t$. However, if $n < m - 1$ or $t' = \varepsilon$, then $\mathbf{clust}_w(u) \neq \mathbf{clust}_w(v)$, because we get an occurrence of u at $i + |t|$. $\square$

The next lemma helps establish the minimum number of non-overlapping occurrences of a word whose cluster is equal to the cluster of a longer word.

Lemma 4. *Let $u^2 \neq v^2$ be two squares in some word w with $u \leq_p v$ and $\mathbf{clust}_w(u) = \mathbf{clust}_w(v)$. If their corresponding rightmost occurrences start at positions u_s and v_s , respectively, then $|u_s - v_s| \geq |u|$.*

Proof. **Case 1:** $v_s > u_s$. Since u occurs at v_s , if $v_s - u_s < |u|$, then u occurs at v_s , between u_s and $u_s + |u|$, and by Lemma 1 u is a repetition. Furthermore, v occurs at $u_s + |u|$, between v_s and $v_s + |v|$ so by Lemma 1 v is also a repetition, which contradicts Lemma 3.

Case 2: $0 < u_s - v_s < |u|$. There are two possibilities to consider, depending on the relative position of $u_s + |u|$ and $v_s + |v|$.

If $u_s + |u| \leq v_s + |v|$, then u occurs inside v other than as a prefix, which leads to u occurring after the last occurrence of v , contradicting $\mathbf{clust}_w(u) = \mathbf{clust}_w(v)$.

If $u_s + |u| > v_s + |v|$, then there is an occurrence of u at position $v_s + |v|$, and an occurrence of v at position u_s , and by Lemma 1 we get that both u and v are non-primitive, which again contradicts Lemma 3. $\square$

Corollary 1. *Let $u_1^2, \dots, u_n^2$ be squares in w such that $\mathbf{clust}_w(u_1) = \dots = \mathbf{clust}_w(u_n)$. Then, $|\mathbf{clust}_w(u_1)| > n$.*

This means that for a chain of square roots $u_1 \leq_p \dots \leq_p u_n$, in both extreme cases, that is, when $|\mathbf{clust}(u_1)| > \dots > |\mathbf{clust}(u_n)|$ or when $|\mathbf{clust}(u_1)| = \dots = |\mathbf{clust}(u_n)|$, our hypothesis for single chains holds, $|\mathbf{clust}(u_1)| > n$. In the next section we prove the hypothesis for single chains in the general case.

3 Bound for single chains

In this section we first prove the upper bound from Conjecture 2 in the special case of single chains. That is, we show that for a set of squares $S = \{u_1^2, \dots, u_n^2\}$ in an arbitrary word w , with $u_1 <_p \dots <_p u_n$, the inequality $|\mathbf{clust}_w(u_1)| > n$ holds. Afterwards we will discuss the challenge in extending the result to multiple chains, the sharpness of the bound and the existence of words $w, u_1, \dots, u_n$ for every possibility of cluster sizes satisfying the bound.

§§ **Upper bound.** For a prefix $x \leq_p u$, we say that the x -representative (x -rep) of u^2 is the longest prefix of u^2 which ends in x . Note that this x -rep is of length at least $|u| + |x|$. Formally, the x -rep of u^2 is $uu'x \leq_p u^2$ such that for every y with $uyx \leq_p u^2$ we have that $|y| \leq |u'|$.

Let w be a word which contains u^2 as a factor. For the first (leftmost) occurrence in w of the x -rep $uu'x$ of square u^2 , let u_s be its starting position and u_m be the start of the u' part, that is, $u_m = u_s + |u|$. We define the x -anchor of u^2 in w as the rightmost occurrence of a factor x in the first occurrence of the x -representative of square u^2 in w . This x -anchor is denoted by $\Psi_w(u^2, x)$. If the x -rep of u^2 is $uu'x$, then $\Psi_w(u^2, x) = u_s + |uu'|$. For example, in the word $w = abaabcabaabab$ we have the square $u = (aba)^2$ starting at position 7. The **a**-rep of u^2 is **abaaba** = u^2 , first occurring at position 7, so $\Psi_w(u^2, a) = 7 + 5 = 12$. The **ab**-rep of u^2 is **abaab**, first occurring at position 1, therefore $\Psi_w(u^2, ab) = 1 + 3 = 4$.

It is one of the novel aspects of our method that we associate a position to a square, its x -anchor, which may not even be close to where the square occurs. Later, in Theorem 2 we reassign some of the squares with colliding x -anchors, further removing the close relation of earlier methods between the counted repetition and the position assigned to it, by which it was counted. This allows us to state a result which - albeit partial - is the first sharp bound on distinct repetitions of any type that we are aware of.

Lemma 5. *Let w be an arbitrary word with two square factors u^2, v^2 such that $u <_p v$, and let $x \leq_p u$ be a common prefix of theirs. If $\Psi_w(u^2, x) = \Psi_w(v^2, x)$, then $u = t^k$ for some primitive word t with $|t| < |x|$ and $k \geq 2$. Moreover, $tu'x \leq_p v$, where $u'x$ is the longest prefix of u bordered by x .*

Proof. Assume $\Psi_w(u^2, x) = \Psi_w(v^2, x)$. We distinguish three cases based on the relative positions of u_s , u_m and v_m , and derive contradictions in all of them, except when u is non-primitive with its root shorter than x . Note that $v_m \leq u_m$ always holds, since $u \leq_p v$ implies $\Psi_w(u^2, x) - u_m \leq \Psi_w(v^2, x) - v_m$.

Fig. 4. Possible cases when $\Psi_w(u^2, x) = \Psi_w(v^2, x)$.

(1) If $v_m \leq u_s$ (Fig. 4(1)), then the x -rep of u^2 is a factor of v , whence it also occurs at $u_s - |v|$, contradicting that the occurrence at u_s was the first.

(2) If $v_m = u_m$ (Fig. 4(2)), then u is a suffix of v and since $|v| > |u|$, we have $v = tu$, for a non-empty word t . Let the x -rep of u^2 be $uu'x$. From $\Psi_w(u^2, x) = \Psi_w(v^2, x)$, we get that the x -rep of v^2 is $vu'x$. However, $tu'x \leq_p v$, so

$$\Psi_w(v^2, x) \geq v_s + |vtu'| > v_s + |vu'| = \Psi_w(v^2, x),$$

a contradiction.

(3) If $u_s < v_m < u_m$, then let the x -reps of u^2 and v^2 be $uu'x$ and $vzu'x$, respectively, where z is the non-empty word starting at v_m and ending at $u_m - 1$. Both $zu'x$ and u are prefixes of v , so if $|zu'x| < |u|$, then $zu'x \leq_p u$, therefore

$$\Psi_w(u^2, x) \geq u_s + |uzu'| > u_s + |uu'| = \Psi_w(u^2, x),$$

a contradiction. If $|zu'x| \geq |u|$, then since $u \leq_p v$, there is an occurrence of u at v_m , so by Lemma 1, we get that u is not primitive (see Fig. 4(3a)).

Now let $u = t^k$, with t primitive and $k \geq 2$. If $|uu'| \geq |v|$ then a conjugate of v is a prefix of uu' , and synchronization, together with $u = t^k \leq_p v$, gives $v = t^m$, where $m > k$. This, in turn, means that u^2 and hence the x -rep of u^2 occurs at position v_s , so the occurrence at u_s is not the leftmost, a contradiction. We are left with the case $|uu'| < |v|$. Given that we have an occurrence of x at u_s , if that

x finishes before position v_m , then there is a copy of x located $|v|$ positions to the right in v^2 , that is, $\Psi_w(v^2, x) \geq u_s + |v| > u_s + |uu'| = \Psi_w(u^2, x)$, contradicting $\Psi_w(v^2, x) = \Psi_w(u^2, x)$.

Hence, we get that $v_m - u_s < |x|$, which means $|t| < |x|$ (see Fig. 4(3b)). As x is a prefix of $u = t^k$, it has the form $x = t^\ell t'$ for some $\ell < k$ and $t' \leq_p t$. The longest prefix of $u = t^k$ bordered by x is $t^{k-1}t' = u'x$. As $u_m > v_m$, we get that $tt^{k-1}t' = tu'x \leq_p v$. $\square$

Corollary 2. *Let $u_1^2, \dots, u_n^2$ and $v_1^2, \dots, v_n^2$ be squares in a word w with their roots from the same chain, and x a common prefix of theirs, such that $\Psi_w(u_i^2, x) = \Psi_w(v_i^2, x)$ and $|u_i| < |v_i|$ for all $i \in \{1, \dots, n\}$. Then, there exists a primitive word t shorter than x , such that $u_i = t^{k_i}$ with $k_i \geq 2$, for all $i \in \{1, \dots, n\}$.*

Proof. From Lemma 5, whenever the x -anchor of some u_i^2 and v_i^2 coincide, there is some primitive t_i with $|t_i| < |x|$ such that $u_i = t_i^{k_i}$ with $k_i \geq 2$ and $t_i x$ is a prefix of v_i . Given that the roots of these squares form a prefix chain, we get that the words $t_i x$ also form a prefix chain, that is, for all $i, j \in \{1, \dots, n\}$ either $t_i x \leq_p t_j x$ or $t_j x \leq_p t_i x$. Furthermore, since x is a common prefix of all the squares, we have $x \leq_p t_i x$, so x has period $|t_i|$, and therefore, trivially, so does $t_i x$. For any pair t_i, t_j , with $|t_i| \leq |t_j|$, we know that $t_i x \leq_p t_j x$, so $t_i x$ also has period $|t_j|$. Since $|t_i x| > |t_i| + |t_j| > |t_i| + |t_j| - \gcd(|t_i|, |t_j|)$, we can apply the theorem of Fine and Wilf and get that t_i and t_j have a common primitive root t . We already know that t_i and t_j are primitive, so $t_i = t_j = t$. $\square$

Theorem 2. *For all words w and squares $u_1^2, \dots, u_n^2$ in w with $u_1 <_p \dots <_p u_n$:*

$$|\text{clust}_w(u_1)| \geq n + 1.$$

Proof. Our strategy consists of assigning a distinct occurrence of u_1 to each u_i^2 , and finding one extra occurrence of u_1 not assigned to any square. Let $x = u_1$.

One by one, in decreasing order of length, we assign to each u_i the position $\Psi_w(u_i^2, x)$, as long as this position has not been previously assigned to a longer square. If all squares have been assigned such a unique position, we are done, because this renders n distinct occurrences of x , while the leftmost occurrence of x in w cannot be the x -anchor of any square.

Otherwise, there is some u^2 in the chain such that $\Psi_w(u^2, x)$ has been assigned to a longer square and we may assume that u^2 is the longest such square. By Lemma 5, $u = t^k$ for some primitive word t and $k \geq 2$. To all squares u_i^2 , if $|u_i| > |u|$ or $u_i \notin t^+$, then assign the position $\Psi_w(u_i^2, x)$. Those are all distinct, by Corollary 2. After this, the roots of all squares which do not have an assigned position yet are powers of t with exponent at most k . Let those squares have roots $t^{k_1}, \dots, t^{k_m}$ with $1 < k_1 < \dots < k_m = k$. Let $x = t^\ell t'$ for some non-empty $t' \leq_p t$. We know that $|x| \leq |t^{k_1}|$, since $x = u_1$, so $k_1 > \ell$, which means $|t^{k_m}| \geq (m + \ell) \cdot |t|$. By Lemma 5, we know that $t^{k_m} t' \leq_p u_n$, so $t^{m+\ell} t' \leq_p u_n$.

Let s_i be the leftmost position where $t^i x$ occurs in w . We assign the position $p_i = s_i + i \cdot |t|$ to the square $(t^{k_i})^2$. It is easy to see that at each of these positions we have an occurrence of x starting, and that $p_i \neq p_j$ whenever $i \neq j$ (in fact,

$p_i < p_{i+1}$). Therefore, what is left to show is that p_i does not coincide with any position assigned in the first phase.

Assume there exists $v \in \{u_1, \dots, u_n\} \setminus \{t^{k_1}, \dots, t^{k_m}\}$ such that $\Psi_w(v^2, x) = p_i = s_i + i \cdot |t|$. From the definition of x -anchor we get that the factor preceding p_i is vv' , where $v'x \leq_p v$. We derive a contradiction in all cases depending on v , which is either (1) a power of t , (2) some other prefix of a power of t or (3) neither, and as such, has ut' as a prefix:

(1) $v = t^q$. The powers of t which have been assigned a position in the first phase are longer than u , so $q \leq k$ is not possible, hence $q > k$. In this case there is a factor t^q preceding $\Psi_w(v^2, x) = p_i$, which means that $t^k t' = t^{m+\ell} t' = t^m x$ occurs at $p_i - q \cdot |t|$, but $t^i x \leq t^m x$ and $p_i - q \cdot |t| < p_i - k \cdot |t| = s_i$, a contradiction.

(2) $v = t^j t''$, with non-empty $t'' <_p t$. In this case, $v' = t^r$ for some $r \geq 0$. If $r \geq i$, then $t^i x$ occurs at $s_i - |v|$, a contradiction. For the other case, let $t = t'' t'''$. If $r < i$, then we get that $p_i - |v'|$ is immediately preceded by $t = t'' t'''$, because of the definition of p_i , and it is also immediately preceded by $t''' t''$, because it is a suffix of v . Hence, $t'' t''' = t''' t''$, so t is non-primitive, a contradiction.

(3) $v = t^k t' z$, for some z with $t' z \not\leq_p t^\omega$. Let the leftmost x -rep of v^2 start at position v_s in w . We have $\Psi_w(v^2, x) \geq v_s + |v| > v_s + |u|$ from the definition of $\Psi_w(v^2, x)$ and the shape of v . We know $v_s + |u| > v_s + m \cdot |t|$, because $u = t^k = t^m t^\ell$ and $\ell \geq 1$. Since $t^i x \leq_p t^k x \leq_p v$ and s_i is the leftmost position where $t^i x$ occurs we get $s_i \leq v_s$, and so $v_s + m \cdot |t| \geq s_i + m \cdot |t|$. Further, $s_i + m \cdot |t| \geq s_i + i \cdot |t|$, because $i \in \{1, \dots, m\}$, and finally, $s_i + i \cdot |t| = p_i$ by definition of p_i . Putting it all together we get $\Psi_w(v^2, x) > p_i$, contradicting the assumption that they coincide.

We have assigned a distinct occurrence to each square u_i . Moreover, the leftmost x in w cannot be the x -anchor of any square and occurs no later than $s_1 (< p_1)$, and so it has not been assigned yet, therefore the theorem holds. $\square$

Now let us see why this result cannot be applied in a straightforward manner to cases when the prefix order is only a partial order on the roots of the squares. Consider the chains $x_1 <_p \dots <_p x_k$, $u_1 <_p \dots <_p u_m$ and $v_1 <_p \dots <_p v_n$, where $x_k <_p u_1$ and $x_k <_p v_1$, but u_1 and v_1 are incomparable by $<_p$ (as in Fig. 3). By Theorem 2 we know that $|\text{clust}(u_1)| \geq m+1$ and $|\text{clust}(v_1)| \geq n+1$, so $|\text{clust}(x_k)| \geq m+n+2$. Unfortunately, for the clusters of x_i , $i < k$, we cannot use the same argument as before, since $\Psi_w(u_j^2, x_i) = \Psi_w(v_\ell^2, x_i)$ is possible without either u_j or v_ℓ being non-primitive, a key condition in the proof of the previous theorem. Take, for example, $u_j = yzzyz$ and $v_\ell = zyz$, for some words y, z , both bordered by x_i , and incomparable by $\leq_p$. Then, in the word $w = yzzyzzyzzyz$ we get $\Psi_w((yzzyz)^2, x_i) = |w| - |x_i| + 1 = \Psi_w((zyz)^2, x_i)$. However, as this example shows, in such a case u_j and v_ℓ have a special structure resembling the reverses of the FS double squares analyzed in [8], so a refinement of the anchor positions and the assignment algorithm might work.

References

1. Bannai, H., I, T., Inenaga, S., Nakashima, Y., Takeda, M., Tsuruta, K.: The "runs" theorem. *SIAM J. Comput.* **46**(5), 1501–1514 (2017)
2. Blanchet-Sadri, F., Mercas, R., Scott, G.: Counting distinct squares in partial words. *Acta Cybern.* **19**(2), 465–477 (2009)
3. Blanchet-Sadri, F., Mercas, R., Scott, G.: A generalization of Thue freeness for partial words. *Theor. Comput. Sci.* **410**(8–10), 793–800 (2009)
4. Crochemore, M., Fazekas, S., Iliopoulos, C., Jayasekera, I.: Number of occurrences of powers in strings. *Internat. J. Found. Comput. Sci.* **21**(4), 535–547 (2010)
5. Crochemore, M., Iliopoulos, C., Kubica, M., Radoszewski, J., Rytter, W., Waleń, T.: The maximal number of cubic runs in a word. *J. Comput. System Sci.* **78**(6), 1828–1836 (2012)
6. Crochemore, M., Rytter, W.: Squares, cubes, and time-space efficient string searching. *Algorithmica* **13**(5), 405–425 (1995)
7. Dekking, F.: On repetitions of blocks in binary sequences. *J. Combin. Theory Ser. A* **20**, 292–299 (1976)
8. Deza, A., Franek, F., Thierry, A.: How many double squares can a string contain? *Discrete Appl. Math.* **180**, 52–69 (2015)
9. Fraenkel, A., Simpson, J.: How many squares must a binary sequence contain? *Electron. J. Combin.* **2**, #R2 (1995)
10. Fraenkel, A., Simpson, J.: How many squares can a string contain? *J. Combin. Theory Ser. A* **82**(1), 112–120 (1998)
11. Gusfield, D.: *Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology*. Cambridge University Press (1997)
12. Hickerson, D.: Less than $2n$ distinct squares in a word of length n (2003), communicated by Dan Gusfield
13. Štěpán Holub: Prefix frequency of lost positions. *Theoretical Computer Science* **684**, 43 – 52 (2017)
14. Ilie, L.: A simple proof that a word of length n has at most $2n$ distinct squares. *J. Combin. Theory Ser. A* **112**(1), 163–164 (2005)
15. Ilie, L.: A note on the number of squares in a word. *Theoret. Comput. Sci.* **380**(3), 373–376 (2007)
16. Jonoska, N., Manea, F., Seki, S.: A stronger square conjecture on binary words. In: *Proc. 40th SOFSEM. LNCS*, vol. 8327, pp. 339–350 (2014)
17. Kociumaka, T., Radoszewski, J., Rytter, W., Waleń, T.: Maximum number of distinct and nonequivalent nonstandard squares in a word. *Theor. Comput. Sci.* **648**(C), 84–95 (Oct 2016)
18. Kolpakov, R., Kucherov, G.: Finding maximal repetitions in a word in linear time. In: *Proc. 40th FOCS*. pp. 596–604. IEEE Computer Society Press (1999)
19. Lothaire, M.: *Combinatorics on Words*. Cambridge University Press (1997)
20. Rampersad, N., Shallit, J., Wang, M.w.: Avoiding large squares in infinite binary words. *Theor. Comput. Sci.* **339**(1), 19–34 (2005)
21. Storer, J.A.: *Data Compression: Methods and Theory*. Comp. Sci. Press, Inc. (1988)
22. Thue, A.: Über unendliche Zeichenreihen. *Kra. Vidensk. Selsk. Skrifter. I Mat. Nat. Kl.* **7** (1906)
23. Thue, A.: Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Kra. Vidensk. Selsk. Skrifter. I Mat. Nat. Kl.* **1** (1912)