Clusters of repetition roots: single chains

Szilárd Zsolt Fazekas[1] and Robert Mercaş[2,3]

[1] Akita University, Department of Computer Science and Engineering, 1-1 Tegata Gakuen-machi, Akita City, 010-8502, Japan, szilard.fazekas@gmail.com
[2] Kiel University, Department of Computer Science, D-24098 Kiel, Germany,
[3] King's College London, Department of Informatics, London WC2R 2LS, UK,
rgm@informatik.uni-kiel.de

Abstract. This paper proposes a new approach towards solving an over 20 years old conjecture regarding the maximum number of distinct squares that a word can contain. We look at clusters of repetitions roots, that is, the set of positions where the root u of a repetition u^ℓ occurs. The theory could provide short and general proofs for bounds on the maximum number of different kinds of repetitions. Here we lay the foundation of this theory and show bounds for the case when the roots of the repetitions form a chain with respect to the prefix order. We also illustrate a simple argument for bounding the number of distinct runs using our approach.

1 Introduction

Repetitions (periodicities) in words are fundamental objects, due primary to their importance in word combinatorics [20] as well as in various applications, such as string matching algorithms [7], molecular biology [14], or text compression [22]. 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. Furthermore, in this case x also represents a period of the considered word. For higher powers we also have cubes, such as xxx, or in general ℓ-repetitions as for $\underbrace{xx \cdot x}_{\ell} = x^\ell$.

A string is said to be square-free or repetition-free if it contains no squares. It was shown by Thue [23, 24] 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 a binary string must contain is three [12].

Since then, there have been extensive studies on testing square-freeness of a string as well as finding squares in a string [7, 1, 8]. However, it is known that the maximum number of squares in a string of length n is $\Theta(n \log n)$. Furthermore, linear time algorithms finding maximal repetitions have also been provided [19].

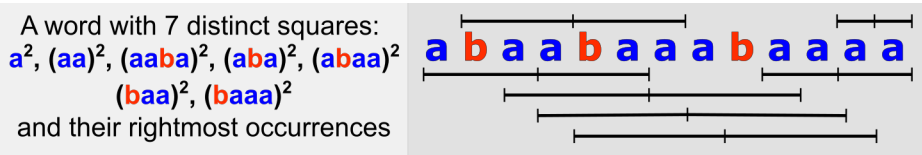
Repetition counting has also been investigated in other settings, as when the length of the root (that is x for an ℓ-repetition x^ℓ) has length as small or large as possible (e.g., see [9, 12, 21]), or for partial words, where words contain extra joker symbols that match every letter of the alphabet (e.g., see [3, 4, 2].

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 runs that a string can contain. A run represents a series of positions in a word that correspond to a maximal extending repetition whose period increases whenever we consider the previous or following letter.

§§ **Problems.** In [13], the authors prove that the maximum number of distinct squares in a word is bounded by twice the length of the word (by looking at the start position of the last occurrence of each square) and conjecture the following:

Conjecture 1. The number of distinct squares in a length n word is less than n.

In the same paper, the authors also provide a construction for words intended as examples of lower bounds for this problem. A simpler construction that is to provide a better lower bound asymptotically was provided only recently in [18], where, considering a binary word with only k occurrences of 1's and even more occurrences of 0's, the authors show the existence of $\frac{2k-1}{2k+2}$ many distinct squares.



Several alternative proofs regarding the $2n$ upper bound were provided, either using combinatorics on words techniques [16], or just calculus [15]. Furthermore, the upper bound was later improved to $2n - \Theta(\log n)$ in [17] by showing that, in addition to the fact that no more than two repetitions can have their last occurrence start at the same position, the number of such positions is bounded.

With this problem in mind, in [5] the authors show that when considering distinct repetitions of a fixed exponent ℓ, the following upper bound is available:

Theorem 2. *For a fixed integer $\ell > 2$, the number of distinct ℓ-powers in a length n word is less than $\frac{n}{\ell-1}$.*

For the case of cubes, a shorter proof was given in [6].

However, the study of this later problem involving repetitions of a higher fixed exponent, has its inspiration in the investigation of the maximum number of runs that a word can have. A run represents a repetition in a string whose period is less than half and which cannot be extended to either left or right in the given word, without breaking the periodicity. The bound on this number was already conjectured to be less than n for every word of such a length since [19]. Only recently [1], with a simple and elegant proof this is showed to be the case.

Theorem 3. *The number of runs in a length n word is less than n.*

Furthermore, in the same paper, the authors also provide the best known upper bound on the sum of exponents of runs in a word.

Theorem 4. *The sum of exponents of runs in a length n word is less than $3n$.*

§§ **Discussion of Techniques.** The technique we use here considers the global properties of occurrences of repetitions in a word, unlike all previous approaches where the bounds were mostly derived form local properties.

The main idea behind the approach is to group the repetitions we want to count by their root. All repetitions whose roots form a chain with respect to $\leq_p$, will be in one group. Then, we show that for every element in a group there are at least a number of positions with are not part of the positions of another element's group. Moreover, when comparing two separate groups, we show that each group has an extra number of positions, rendering us with enough options as to take care of their common prefix. From there it follows directly that there are less elements in all groups combined than positions in the string, giving us all the results presented later.

§§ **Our Results.** In this work we prove Conjecture 1 to be true, as a direct consequence of an alternative proof of Theorem 2. As a byproduct, we show that, surprisingly, this problems can be connected with the count of the number of runs in a word, and provide alternative proofs for both Theorems 3 and 4.

It is worth noting that while most of the results concerning the bounds on the maximum number of distinct squares were obtained using local properties of strings, 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. However, these were never connected.

In this work, we show the first connection between all the results by looking only at prefixes of such words, and the set of positions where these appear.

§§ **Preliminaries.** We end this section with a few definitions we use.

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 prefix/suffix/factor is *proper* if it is non-empty and not equal to w. We call p a *period* of w if every p apart positions in w are the same. The *minimal period* is given by the smallest such p.

A *repetition* represents consecutive catenations of the same word. In particular, a *square* is given by two such consecutive repetitions, a *cube* by three, while, in general, an *ℓ-power* (*ℓ-repetition*) represents ℓ such repetitions of the same factor. A *run* is given by the positions in the word that contain a maximal repetitive factor with a period at most as long as the length of the factor (a repetition is maximal, if taking a previous or following position, breaks the repetition).

If a word is not a repetition, then it is called *primitive*. Furthermore, if $w = u^\ell$ is an ℓ-repetition we say that u is a *root* of w, and for any primitive word t such

that $u = t^k$, we call t the *primitive root* of w. Finally, by t^ω we denote the infinite word consisting in an infinite number of repetitions of t.

To simplify the exposition, for a word we build a *suffix array* structure consisting of all its suffixes, in lexicographical order; that is, once all suffixes are ordered, the 5th suffix in lexicographical order occurs in position 5 in the array.

2 Clusters of repetition roots

In this section we introduce the clusters of repetition roots and we prove some fundamental properties of these clusters and their relations with other clusters.

In the following we work a lot with sets of words based on their common prefix. First we make the following trivial remarks:

Observation 5 *The set of suffixes of a word sharing a common prefix are contiguous in the suffix array forming a cluster.*

Observation 6 *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.*

Let us fix some notation. We refer, in general, to a word w containing some ℓ-repetitions. We associate it a suffix array S with S_i the ith suffix of w in lexicographical order, and denote by clust(u), for each factor u^ℓ of w, the *cluster* in S that contains all suffixes having u as a prefix. Figure 1 provides a visual image of how these clusters could be perceived (in the case of $\ell = 2$), arranged one on top of the other.

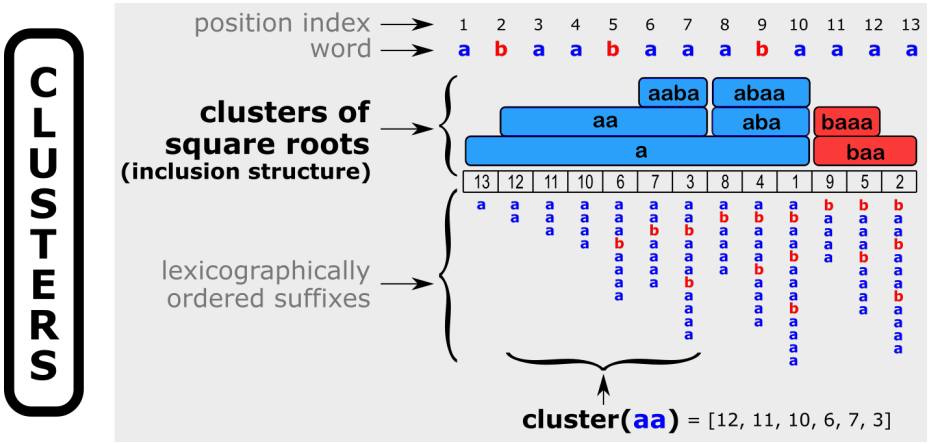


Fig. 1.

As every word, and therefore every suffix, having v as prefix, also has u as prefix, whenever u is a prefix of v, the next observations are easy:

Lemma 7. *For any two clusters* $\mathrm{clust}(u) = [S_{i_1}..S_{j_1}]$ *and* $\mathrm{clust}(v) = [S_{i_2}..S_{j_2}]$ *(with* $i_1 \le j_1$ *and* $i_2 \le j_2$*), the following are immediate:*

1. *if* $u \le_p v$*, then* $\mathrm{clust}(v) \subseteq \mathrm{clust}(u)$*, and vice versa;*
2. *if* u *and* v *are incomparable with respect to* $\le_p$*, then* $j_1 < i_2$ *or* $j_2 < i_1$*, that is, the clusters do not overlap;*
3. *if* $\mathrm{clust}(u) \cap \mathrm{clust}(v) \neq \emptyset$*, then either* $u <_p v$ *or* $v <_p u$*.*

Proof. For item (1) we note that, if u is a prefix of v, then obviously u occurs as a prefix of each element in $\mathrm{clust}(v)$, and the result follows. If on the other hand $\mathrm{clust}(v) \subseteq \mathrm{clust}(u)$, then obviously they must have a common prefix. However, since all elements within $\mathrm{clust}(u)$ have u as a prefix, the result follows again.

Item (2) says that if there exists $i \in \mathrm{clust}(u) \cap \mathrm{clust}(v)$, meaning both u and v appear starting at the same position, then one of them is a prefix of the other.

Item (3) says that if $\mathrm{clust}(u) \cap \mathrm{clust}(v) \neq \emptyset$, then all their elements share u as common prefix. This is direct from (1) and (2). □

Our goal is to show that for ℓ-repetitions $u_1^\ell, \ldots, u_n^\ell$ with a common prefix x, we have $m < \frac{1}{\ell-1} \cdot |\mathrm{clust}(x)|$. In this paper we approach the problem by analyzing the case where $\ell = 2$ and $u_1 \le_p \cdots \le_p u_n$. We will call such a collection of squares, a *(prefix) chain*. Figure 2 illustrates our aim in the case of single chains and more generally for multiple chains of squares. One can see that if our hypothesis for multiple chains holds with $C = 1$, then Conjecture 1 is true, that is, there are fewer distinct squares than the length of the word (note that the maximum number of elements in a cluster is given by the length of the whole word producing the respective suffix array).

The next lemmas explore the situation when there are two clusters which are equal. This is a crucial issue, because if the clusters are all different in length, then the bound trivially follows. Most of the following results can be proved in general for ℓ-repetitions, but in order to simplify the argument, here we usually restrict ourselves to $\ell = 2$. First recall the following well-known result about primitivity of words.

Lemma 8. *[20] A word* w *is primitive if and only if it does not have a third occurrence in* ww*, neither as a prefix, nor as a suffix.*

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 9. *Consider factors* $u^\ell \neq v^\ell$ *with* $\ell > 1$ *of some word* w*, such that* $\mathrm{clust}(u) = \mathrm{clust}(v)$ *and* $u <_p v$*. Then the following hold:*

1. *if* $u = t^m$ *for some primitive* t *and* $m > 1$*, then* $v = t^m t'$ *for some non-empty* $t' \le_p t$*.*
2. *if* $v = t^m$ *for some primitive* t *and* $m > 1$*, then* $u = t^{m-1}t'$ *for some non-empty* $t' \le_p t$*.*

Proof. Trivially, for two words to have equal clusters, their rightmost occurrences must start at the same position.Let the rightmost occurrence of u and v be at position i.

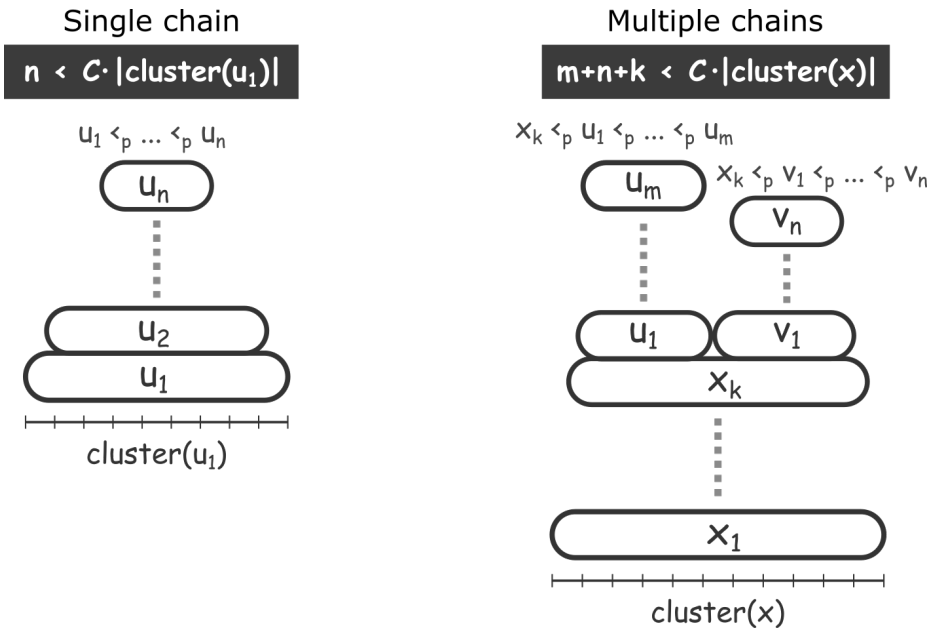


Fig. 2. Framed in black are the inequalities that we conjectured to hold for clusters of repetition roots. Here we prove the one for single chains.

1. Suppose that $u = t^m$ for some primitive t and $m > 1$. Since $\mathrm{clust}(u) = \mathrm{clust}(v)$, we have an occurrence of v wherever u^2 occurs. If $|v| \geq 2|u|$, 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.
2. Suppose that $v = t^m$ for some primitive t and $m > 1$. Since $u \leq_p v$, we get that $u = t^n t'$ for some $n < m$ and $t' \leq_p t$. If $n < m - 1$ or $t' = \lambda$, then $\mathrm{clust}(u) \neq \mathrm{clust}(v)$, because we get an occurrence of u at $i + |t|$. Hence, $u = t^{m-1}t'$ with t' non-empty.

□

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

Lemma 10. *For two squares $u^2 \neq v^2$ with $u \leq_p v$, if $\mathrm{clust}(u) = \mathrm{clust}(v)$, then the rightmost occurrence of u^2 and v^2 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 $|u^2| > |v|$, otherwise u^2 would occur later, at $i + |v|$. We have occurrences of u at i, $i + |u|$ and $i + |v|$. If $\mathrm{clust}(u) = \mathrm{clust}(v)$, then v also occurs at $i + |u|$ and by Lemma 8, we get that v is non-primitive, and by the theorem of Fine and Wilf, the primitive root of u and v is 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. □

Lemma 11. *Let $u^2 \neq v^2$ be two squares with $u \leq_p v$ and $\mathrm{clust}(u) = \mathrm{clust}(v)$, whose rightmost occurrences start at positions u_s and v_s, respectively. Then, either $|u_s - v_s| \geq |u|$ or there exist a primitive word t, an integer $m > 1$ and a non-empty prefix t' of t, such that $u = t^m$ and $v = t^m t'$.*

Proof. If $v_s - u_s \geq |u|$ or $u_s - v_s \geq |u|$, then we are done. Otherwise, by Lemma 10, we know that $u_s \neq v_s$. First, assume $u_s < v_s$. If $v_s - u_s < |u|$, then since u occurs at v_s, by Lemma 8 we get that u is non-primitive. We can apply Lemma 9 and get that $u = t^m$ and $v = t^m t'$ for some primitive t, $m > 1$ and non-empty $t' \leq_p t$.

If $v_s < u_s$ then there are two cases depending on the relative position of $u_s + |u|$ of $v_s + |v|$.

1. $u_s + |u| < v_s + |v|$: this means that there is an occurrence of u inside v, which leads to u occurring at $i + |u_s - v_s|$, a contradiction.
2. $u_s + |u| > v_s + |v|$: since there is an occurrence of u at position $v_s + |v|$, by Lemma 8 we get that u is non-primitive and applying Lemma 9 gives $u = t^m$ and $v = t^m t'$ for some primitive t, $m > 1$ and non-empty $t' \leq_p t$.

□

As a consequence of Lemma 10, we get the following.

Corollary 12. *Let $u_1^2, \ldots, u_n^2$ be squares such that $\mathrm{clust}(u_1) = \cdots = \mathrm{clust}(u_n)$. Then, $|\mathrm{clust}(u_1)| > n$.*

This means that for a chain of square roots $u_1 \leq_p \cdots \leq_p u_n$, in both extreme cases, that is, when $|\mathrm{clust}(u_1)| > \cdots > |\mathrm{clust}(u_n)|$ or when $|\mathrm{clust}(u_1)| = \cdots = |\mathrm{clust}(u_n)|$, our hypothesis for single chains holds, $|\mathrm{clust}(u_1)| > n$.

In the next section we will prove the hypothesis for single chains in the general case. Before that, to end this section, we will illustrate a different type of argument which allows us to put bounds on the number of clusters in a so called maximal chain.

A *maximal chain* is such that for every root u of a square uu occurring in w, if u is comparable by $\leq_p$ with the elements of S, then $u \in S$. For some $u_i, u_j \in S$, we call u_i^2 *covered* by u_j if $u_i \leq_p u_j \leq_p u_i^2$ or $u_i^2 \leq_p u_j$ (see Fig. **??**). The *shortest square root* of a maximal chain S is denoted by $\mathrm{ssr}(S)$. Note that $\mathrm{ssr}(S)$ does not have a border, that is, a prefix which is also a suffix. If it had, say $\mathrm{ssr}(S) = pqp$, for some non-empty word p, then $\mathrm{ssr}(S)^2 = pqppqp$, which contains p^2, so $p \in S$, which would contradict $\mathrm{ssr}(S)$ being the shortest element in S. In what follows, $|v|_x$ denotes the number of times x occurs as a factor of v, in other words, the number of v's suffixes which start with x.

Lemma 13. *For a square series S as above, let m be the number of covered squares in S and let $x = \mathrm{ssr}(S)$. Then, $|u_n|_x \geq m$.*

Proof. If a square u_i^2 is covered by some u_j, then x occurs at position $|u_i|$ in u_j and in all u_k, with $k > j$, because $u_j \leq_p u_k$. In fact, x occurs at position $|u_i|$ even in u_ℓ, for $i < \ell < j$, as $u_i \leq_p u_\ell \leq_p u_j$. □

Lemma 14. *For a maximal chain of square roots $S = \{u_1, \ldots, u_n\}$, with $x = \mathrm{ssr}(S)$, we have*

$$\textbf{\textit{diff}}(x, u_i) + |u_i|_x < \textbf{\textit{diff}}(x, u_{i+1}) + |u_{i+1}|_x,$$

where $\textbf{\textit{diff}}(x, u_i) = |\mathrm{clust}(x)| - |\mathrm{clust}(u_i)|$.

Proof. Since $u_i \leq_p u_{i+1}$, it follows that $|u_i|_x \leq |u_{i+1}|_x$, while $\mathrm{clust}(u_{i+1}) \subseteq \mathrm{clust}(u_i)$ gives $\mathbf{diff}(x, u_i) \leq \mathbf{diff}(x, u_{i+1})$, so both terms in the sum are non-decreasing. In fact, at least one of them increases in each step: if u_{i+1} covers u_i^2, then $|u_i|_x < |u_{i+1}|_x$. If u_{i+1} does not cover u_i^2, then $|\mathrm{clust}(u_{i+1})| < |\mathrm{clust}(u_i)|$), so $\mathbf{diff}(x, u_i) < \mathbf{diff}(x, u_{i+1})$. □

Corollary 15. *For a maximal chain of square roots $S = \{u_1, \ldots, u_n\}$, with $x = \mathrm{ssr}(S)$, we have*

Lemma 16. *For a maximal chain of square roots $S = \{u_1, \ldots, u_n\}$, with $x = \mathrm{ssr}(S)$, we have*

$$|u_n|_x \leq \frac{\textbf{\textit{diff}}(x, u_n)}{2}.$$

Proof. If u_n is primitive, then $|\mathrm{clust}(x)| \geq 2|u_n|_x + |\mathrm{clust}(u_n)| - 2$, because u_n^2 is a factor and two occurrences of u_n cannot overlap in all occurrences of x, therefore apart from the x's occurring in u_n^2 we have an extra x for the other occurrences of u_n.

If $u_n = v^k$, $k > 1$, then $|u_n|_x \geq k$, since v^2 is a factor, and $|v| \geq |x|$ since $x = \mathrm{ssr}(S)$. Moreover, $v^i v^i$ is a factor for all integers $1 \leq i < k$, and $|\mathrm{clust}(v^{k-i})| \geq |\mathrm{clust}(v^k)| + 2i$. The statement follows. □

Corollary 17. *For a maximal chain of square roots $S = \{u_1, \ldots, u_n\}$, with $x = \mathrm{ssr}(S)$, we have*

$$n < \frac{3}{2} \cdot |\mathrm{clust}(x)|.$$

Proof. $\mathbf{diff}(x, x) = 0$ and $|u_1|_x = 1$, since $u_1 = x$. By Lemma 14, for each $i \in \{1, \ldots, n\}$, the sum $\mathbf{diff}(u_i, x) + |u_i|_x$ is strictly increasing, so

$$n < \mathbf{diff}(x, u_n) + |u_n|_x.$$

Since $|\mathrm{clust}(u_n)| \geq 2$, we have $\mathbf{diff}(u_n, x) = |\mathrm{clust}(x)| - |\mathrm{clust}(u_n)| < |\mathrm{clust}(x)| - 1$. From here, applying Lemma 16 gives us the statement. □

By the above we have that for a chain S, the number of clusters is bounded by $3n/2$, where n is the size of the cluster of $\mathrm{ssr}(S)$. This can be improved to $4n/3$ applying Lemma 11. Let us look at the topmost level where two clusters

are equal, that is, suppose $\mathrm{clust}(u) = \mathrm{clust}(v)$ for $u <_p v$ and for all w with $v <_p w$ there exists no z with $\mathrm{clust}(w) = \mathrm{clust}(z)$.

Since $\mathrm{clust}(u) = \mathrm{clust}(v)$, by Lemma 11 we have that $\mathrm{clust}(u) \geq 3$, moreover, there are at least three non-overlapping occurrences of u. From here we get that $|u|_{\mathrm{ssr}} \leq \frac{\mathrm{clust(ssr)}}{3}$, but the consecutive clusters above v are never equal, hence the number of clusters is at most clust(ssr) plus the number of times when two consecutive cluster are equal. The latter is at most $|u|_{\mathrm{ssr}}$, hence we get that the number of clusters is at most $\frac{4 \cdot \mathrm{clust(ssr)}}{3}$.

3 Bound n for single chains

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 $uzx \leq_p u^2$ such that for all y, if $uyx \leq_p u^2$ then $|y| \leq |z|$.

For the first (leftmost) occurrence of the x-rep of square uu, we denote its starting position by u_{xs} and its middle by u_{xm}, that is, $u_{xm} = u_{xs} + |u|$. When x is fixed, we will just refer to the positions as u_s and u_m.

The *x-anchor* of u^2, that is, the rightmost occurrence of a factor x in the first occurrence of the x-representative of square u^2 will be denoted by $\mathbf{first}(u^2, x)$ that is, if the x-rep of u^2 is $uu'x$, then $\mathbf{first}(u^2, x) = u_{xs} + |uu'|$.

Lemma 18. *Let u^2, v^2 be two squares with $u \leq_p v$, and let $x \leq_p u$ be a common prefix of theirs. If $\mathbf{first}(u^2, x) = \mathbf{first}(v^2, x)$ then $u = t^k$ for some primitive word t with $|t| < |x|$ and $k \geq 2$.*

Proof. Assume $\mathbf{first}(u^2, x) = \mathbf{first}(v^2, x)$. 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 $u \leq_p v$.

1. $v_m \leq u_s$. In this case the x-rep of u^2 is a factor of v, therefore it also occurs at $u_s - |v|$, a contradiction.
2. $v_m = u_m$. This means that u is a suffix of v and since $|v| > |u|$, we have $v = tu$, for some non-empty word t. Let the x-rep of u^2 be uzx. From $\mathbf{first}(u^2, x) = \mathbf{first}(v^2, x)$, we get that the x-rep of v^2 is vzx. However, $tzx \leq_p v$, so
$$\mathbf{first}(v^2, x) \geq v_s + |vtz| > v_s + |vz| = \mathbf{first}(v^2, x),$$
a contradiction.
3. $u_s < v_m < u_m$. Let the x-reps of u^2 and v^2 be $uu'x$ and $vzu'x$, respectively, where z is the non-empty suffix of u starting at v_m. If $|zu'x| < |u|$, then $zu'x \leq_p u$, so
$$\mathbf{first}(u^2, x) = u_s + |uzu'| > u_s + |uu'| = \mathbf{first}(u^2, x),$$

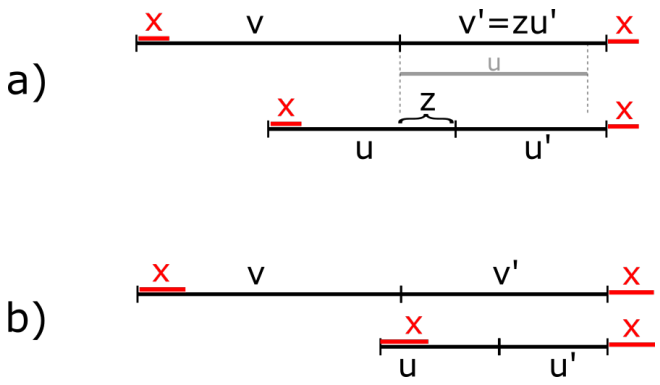


Fig. 3.

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 8, we get that u is not primitive (see Fig. 3(a)). Now let $u = t^k$, with t primitive and $k \geq 2$. If $|uu'| \geq |v|$ then a conjugate of v is a factor 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 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 $\mathbf{first}(u^2, x) \geq u_s + |v| > u_s + |uu'| = \mathbf{first}(u^2, x)$, a contradiction. Hence, we get that $v_m - u_s < |x|$, which means $|t| < |x|$.

□

From the previous lemma, whenever the x-anchor of u^2 and v^2 coincide, we can write $v = t^m z t^n$ and $u = t^k$ with $n < k < m$, $|t| < |x|$, and z is such that x does not occur in v after position $(n-1) \cdot |t|$. As $x \leq_p u$, we get that x is of the form $t^r t'$, for some $r > 0$ and $t' \leq_p t$. Let the x-anchor of u^2 be at position $u_s + i \cdot |t|$, for some $k < i < 2k$. Given the form of x, we know that x occurs also at position $u_s + (i-1) \cdot |t|$, and we claim that this position does not coincide with the x-anchor of any square in the chain.

4 Conclusion

An important observation is that the notion of suffix array is in fact not a necessary one, but its use makes the whole exposition easier to follow. Furthermore, the used notion of clusters has been recently used in several other papers [10, 11] and proves to be of the upmost interest.

Also, we stress the fact that our calculus are not necessarily precise, but are enough to obtain the existing or conjectured upper bounds of our considered problem. A stricter analysis should take into account that two clusters included one in the other, have most of the times different lengths, with strict restrictions regarding their form, otherwise (see Lemma 9). Furthermore, the length of the

longest root within each of these clusters is bounded by half the length of the word (in the case of squares and runs), which provides extra restrictions. Therefore, it is our belief that a more involved analysis of these structures will in fact improve the existing upper bunds by a factor of at least $\log(n)$, where n is the length of the word. A hint in this direction, is also given by the result from [17] where such a factor is shaved off.

Further, we assume that the proof technique presented here is going to find immediate applications in relation to other results, such as identifying partial or hidden repetitions. It is our opinion that this technique might also provide an alternative proof for the count of the total number of runs in a word.

As a final remark, we note that the words constructed in [13, 18] with the scope of providing lower bounds can be further tweaked as to improve the bound.

Indeed, in [13], the authors show that for a construction of a word of the form $\Pi_{i=1}^{m}(0^{i-1}10^i10^{i+1}1)$, the number of distinct squares it contains is very close to $\frac{3}{2}m^2 + 4m - 3 + \frac{odd(i)}{2}$. In the construction from [18], the authors provide a word of the form $\Pi_{i=1}^{m}(0^i1)$ where the lower bound on the number of square, asymptotically goes to $n \cdot \frac{2i-1}{2i+2}$. However, note that both these constructions are one sided. Therefore, considering a complementary inverse catenation to the left, will improve both lower bounds by at least a constant factor, when considering whichever length. These simple catenations to the left provide us with the words

$$\Pi_{i=1}^{m}(01^{m-i+2}01^{m-i+1}01^{m-i})\Pi_{i=1}^{m}(0^{i-1}10^i10^{i+1}1)$$

where the cube $(01)^3$ in the middle can be easily avoided, and the word

$$\Pi_{i=1}^{m}(01^{m-i+1})\Pi_{i=1}^{m}(0^i1).$$

Both constructions increase asymptotically the ratio between the number of distinct squares and the length of the word in the binary case.

5 Acknowledgement

We would like to thank M. Crochemore for helpful discussions on the subject. The work of Robert Mercaş was supported by . The work of Szilárd Zsolt Fazekas was supported by Grant-in-Aid for Scientific Research no. .

References

1. H. Bannai, T. I, S. Inenaga, Y. Nakashima, M. Takeda, and K. Tsuruta. The "runs" theorem. *CoRR*, abs/1406.0263v4, 2014.
2. F. Blanchet-Sadri, I. Choi, and R. Mercaş. Avoiding large squares in partial words. *Theor. Comput. Sci.*, 412(29):3752–3758, 2011.
3. F. Blanchet-Sadri, R. Mercaş, and G. Scott. Counting distinct squares in partial words. *Acta Cybern.*, 19(2):465–477, 2009.
4. F. Blanchet-Sadri, R. Mercaş, and G. Scott. A generalization of Thue freeness for partial words. *Theor. Comput. Sci.*, 410(8-10):793–800, 2009.

5. M. Crochemore, S.Z. Fazekas, C.S. Iliopoulos, and I. Jayasekera. Number of occurrences of powers in strings. *Internat. J. Found. Comput. Sci.*, 21(4):535–547, 2010.
6. M. Crochemore, C.S. Iliopoulos, M. Kubica, J. Radoszewski, W. Rytter, and T. Waleń. The maximal number of cubic runs in a word. *J. Comput. System Sci.*, 78(6):1828–1836, 2012.
7. M. Crochemore and W. Rytter. Squares, cubes, and time-space efficient string searching. *Algorithmica*, 13(5):405–425, 1995.
8. M. Crochemore and W. Rytter. *Jewels of Stringology*. World Academic, Singapore, 2002.
9. F.M. Dekking. On repetitions of blocks in binary sequences,. *J. Combin. Theory Ser. A*, 20:292–299, 1976.
10. T. Ehlers, F. Manea, R. Mercas, and D. Nowotka. k-abelian pattern matching. In *Proc. 18th DLT*, volume 8633 of *LNCS*, pages 178–190, 2014.
11. H. Fernau, F. Manea, R. Mercaş, and M.L. Schmid. Pattern matching with variables: Fast algorithms and new hardness results. In *Proc. 32nd STACS*, LIPIcs, 2015. To appear.
12. A.S. Fraenkel and J. Simpson. How many squares must a binary sequence contain? *Electron. J. Combin.*, 2:#R2, 1995.
13. A.S. Fraenkel and J. Simpson. How many squares can a string contain? *J. Combin. Theory Ser. A*, 82(1):112–120, 1998.
14. Dan Gusfield. *Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology*. Cambridge University Press, 1997.
15. D. Hickerson. Less than $2n$ distinct squares in a word of length n, 2003. communicated by Dan Gusfield.
16. L. Ilie. 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.
17. L. Ilie. A note on the number of squares in a word. *Theoret. Comput. Sci.*, 380(3):373–376, 2007.
18. Natasa Jonoska, F. Manea, and Shinnosuke Seki. A stronger square conjecture on binary words. In *Proc. 40th SOFSEM*, volume 8327 of *LNCS*, pages 339–350, 2014.
19. R. Kolpakov and G. Kucherov. Finding maximal repetitions in a word in linear time. In *Proc. 40th FOCS*, pages 596–604. IEEE Computer Society Press, 1999.
20. M. Lothaire. *Combinatorics on Words*. Cambridge University Press, 1997.
21. Narad Rampersad, Jeffrey Shallit, and Ming-wei Wang. Avoiding large squares in infinite binary words. *Theor. Comput. Sci.*, 339(1):19–34, 2005.
22. James A. Storer. *Data Compression: Methods and Theory*. Computer Science Press, Inc., 1988.
23. A. Thue. Über unendliche Zeichenreihen. *Kra. Vidensk. Selsk. Skrifter. I Mat. Nat. Kl.*, 7, 1906.
24. A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Kra. Vidensk. Selsk. Skrifter. I Mat. Nat. Kl.*, 1, 1912.