

The Maximum Number of Squares in a Tree

Maxime Crochemore^{1,3}, Costas S. Iliopoulos^{1,4}, Tomasz Kociumaka², Marcin Kubica², Jakub Radoszewski^{2*}, Wojciech Rytter^{2,5**}, Wojciech Tyczyński², and Tomasz Walen^{2,6}

¹ Dept. of Informatics, King's College London, London WC2R 2LS, UK
[maxime.crochemore,csi]@dcs.kcl.ac.uk

² Faculty of Mathematics, Informatics and Mechanics,
University of Warsaw, Warsaw, Poland

[kociumaka,kubica,jrad,rytter,w.tyczynski,walen]@mimuw.edu.pl

³ Université Paris-Est, France

⁴ Faculty of Engineering, Computing and Mathematics,
University of Western Australia, Perth WA 6009, Australia

⁵ Faculty of Mathematics and Computer Science,
Copernicus University, Toruń, Poland

⁶ Laboratory of Bioinformatics and Protein Engineering,
International Institute of Molecular and Cell Biology in Warsaw, Poland

Abstract. We show that the maximum number of different square substrings in unrooted labelled trees behaves much differently than in words. A substring in a tree corresponds (as its value) to a simple path. Let $\text{sq}(n)$ be the maximum number of different square substrings in a tree of size n . We show that asymptotically $\text{sq}(n)$ is strictly between linear and quadratic orders, for some constants $c_1, c_2 > 0$ we obtain:

$$c_1 n^{4/3} \leq \text{sq}(n) \leq c_2 n^{4/3}.$$

1 Introduction

Repetitions are a fundamental notion in combinatorics and algorithmics on words. The basic type of a repetition are squares: words of the type zz , where $z \neq \varepsilon$. (By ε we denote the empty word.) In this paper we consider square substrings corresponding to simple paths in labelled trees. If a tree is a single path then it is a problem of classical repetitions in strings. Combinatorics of squares in classical strings has been investigated in [7,9,10] and for partial words in [3]. Squares were also studied in the context of games, e.g. in [8].

Repetitions in trees and graphs have already been considered, for example in [4,1,2]. The number of square substrings in general graphs dramatically increases — it can be exponential, even in case of binary alphabet.

Assume we have a tree T whose edges are labelled with symbols from an alphabet Σ . By $|T|$ we denote the size of the tree, that is the number of nodes.

* The author is supported by grant no. N206 568540 of the National Science Centre.

** The author is supported by grant no. N206 566740 of the National Science Centre.

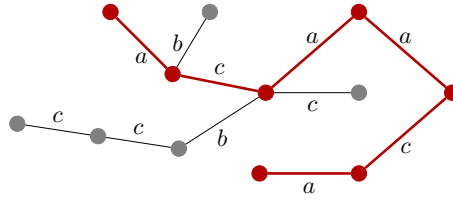


Fig. 1. There are 4 square substrings in this tree: aa , $acaaca$, $bbcc$, cc . Note that cc occurs twice. The longest is $acaaca$ and it corresponds to a path marked with a solid line in the figure.

If u and v are two nodes of T , then by $val(u, v)$ we denote the sequence of labels of edges on the path from u to v . We call $val(u, v)$ a substring of T . (Note that a substring is a string, not a path.) Figure 1 illustrates a square substring in a sample tree. We consider only simple paths: this means that vertices of a path do not repeat (though edges on the path do not need to have distinct labels).

For a tree T , by $sq(T)$ we denote the number of different square substrings in T . For the tree T from Fig. 1, we have $sq(T) = 4$. Let $sq(n)$ be the maximum of $sq(T)$ over all trees of size n . We show that $sq(n) = \Theta(n^{4/3})$. Thus $sq(n)$ has different asymptotics than the maximum number of different square substrings in a standard word (a single path tree) of length n , which is known to be $\Theta(n)$ [7].

We introduce a family of trees which we call combs. The lower bound for $sq(n)$ turns out to be realized by trees from this family, and such trees also play an important role in the proof of the upper bound. Before we show the general upper bound, we provide some intuition behind this proof by showing the same upper bound for combs and for *special squares* of the form $(a^i b a^j)^2$ in general trees.

2 Bounds for Combs

Before we show a general $O(n^{4/3})$ bound on the number of squares in a tree, we analyze the number of squares for a family of trees which we call *standard combs*. The notion of combs is generalized later in the paper.

A *standard comb* is a labelled tree that consists of a path called the *spine*, with at most one *branch* attached to each node of the spine. All spine-edges are labelled with the letter a . Each branch is a path starting with the letter b , followed by a number of a -labelled edges, see Fig. 2.

As we show in the theorem below, there exists a family T_m of standard combs for which $sq(T_m) = \Omega(|T_m|^{4/3})$. From this one easily obtains $sq(n) = \Omega(n^{4/3})$ for any n . In this section we also prove an upper bound of $O(n^{4/3})$ for the number of squares in a standard comb of size n . This proof is extensively used throughout the proof of the same upper bound for general trees, given in the following sections. Hence, our family of standard combs T_m meets the asymptotic upper bound for $sq(n)$ for general trees.

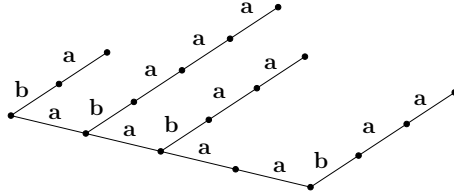


Fig. 2. A standard comb containing 11 square substrings.

For $m = k^2$ we define a set $Z_m = \{1, \dots, k\} \cup \{i \cdot k : 1 \leq i \leq k\}$. For example, if $m = 9$, then $Z_m = \{1, 2, 3, 6, 9\}$.

Lemma 1. *Assume m is a square of a positive integer. Then for each $0 < j < m$ there exist $u, v \in Z_m$ such that $u - v = j$.*

Proof. Each number $0 < j < m$ can be written as $p\sqrt{m} - q$, where $0 < p, q \leq \sqrt{m}$. This formula corresponds to distance between points q and $p\sqrt{m}$. \square

For $m = k^2$ we define a standard comb T_m as follows: T_m consists of a spine of length m with vertices numbered from 1 to m , and branches of the form ba^m attached to each vertex $j \in Z_m$ of the spine, see Fig. 3.

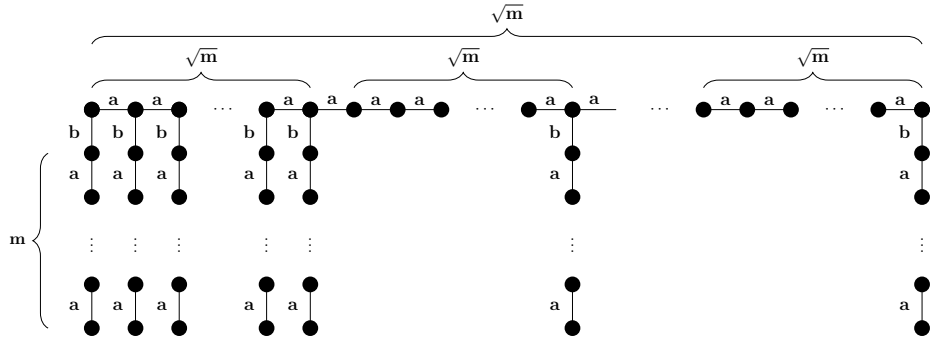


Fig. 3. The structure of a standard comb T_m .

Theorem 1. [Lower Bound Theorem]

For each tree T_m we have $sq(T_m) = \Omega(|T_m|^{4/3})$.

Proof. From Lemma 1, for every $0 < j < m$ there are two nodes u, v of degree 3 on the spine with $distance(u, v) = j$. Hence, T_m contains all squares of the form $(a^i ba^{j-i})^2$ for $0 \leq i \leq j$ and $0 < j < m$. Altogether this gives $\Omega(m^2)$ different squares. Note that $|T_m| = O(m\sqrt{m})$. Hence, the number of square substrings in T_m is $\Omega(|T_m|^{4/3})$. \square

Lemma 2. *The number of squares in a standard comb of size n is $O(n^{4/3})$.*

Proof. Let T be a standard comb of size n . Note that T contains only square substrings of the form $(a^i)^2$ or $(a^i b a^j)^2$. The number of squares of the former type is $O(n)$. We need to bound the number of squares of the latter type (special squares). Any occurrence of a special square starts and ends within two different branches of T .

There are at most $n^{4/3}$ different special squares for which $i < n^{2/3}$ and $j < n^{2/3}$. Hence, it suffices to prove that there are $O(n^{4/3})$ special square substrings of T for $i \geq n^{2/3}$ or $j \geq n^{2/3}$, we call such special squares *long*.

A branch of a standard comb is called *long* if it contains at least $n^{2/3}$ nodes. Note that there are $O(n^{1/3})$ long branches in T . Any occurrence of a long special square has at least one endpoint in a long branch.

Consider a node u located in a branch B of T and a long branch B' . There is at most one occurrence of a long special square that starts in u and ends within the branch B' . Indeed, if there are i a -labelled edges on the path from u to the spine and k edges on the path connecting the branches B and B' then the considered square $(a^i b a^{k-i})^2$ uniquely determines its other endpoint. Hence, the total number of long special squares is bounded by the number of nodes u multiplied by the number of long branches B' , that is, by $O(n^{4/3})$. This completes the proof. \square

3 Prelude to Upper Bound Proof

In this section we show a tight upper bound for *special squares*, defined at the end of Section 1. Along the way we introduce some part of the machinery for the general proof. Define a *double tree* $\mathcal{D} = (T_1, T_2, R)$ as a labelled tree consisting of two disjoint (except one vertex) trees T_1, T_2 with a common root R . The size of \mathcal{D} is defined as $|\mathcal{D}| = |T_1| + |T_2| - 1$. The substrings of \mathcal{D} are defined as values of paths which start within T_1 and end in T_2 . An example of a double tree is shown in Fig. 4, T_1 lies below R (lower tree) while T_2 above R (upper tree).

A directed rooted labelled tree is *deterministic* if the edges going down from the same vertex have different labels. Note that a tree is deterministic if and only if it is a trie (also called a prefix tree) of the values of the paths from R to the leaves. A double tree is *deterministic* if each of the trees T_1, T_2 treated as a directed tree with root R is deterministic. A double deterministic tree is also called here a *D-tree*. The double tree in Fig. 4 is a sample D-tree.

Lemma 3. *For each double (possibly nondeterministic) tree there exists a D-tree with at most the same number of vertices and the same set of substrings (going from T_1 to T_2).*

Proof. For a moment let us direct each tree T_i down from R (treated as a root). Assume we have a vertex v with edges $(v, u), (v, w)$ going to its children and labelled with the same letter a . Then we can *glue* the vertices u, w . We can perform such operation going top-down from the root in a BFS traversal. Note

that the resulting trees T_i are deterministic, their sizes could only decrease, and the set of the substrings of the D-tree remains unchanged. \square

A path in a tree T is said to be anchored in a node $R \in T$ if R lies on this path. A square is *anchored* in R if it is a value of a path anchored in R .

A path p from v to u in a D-tree is called a *D-square* if $v \in T_1, u \in T_2, \text{val}(v, u)$ is a square and its midpoint lies within T_1 , and amongst all such paths of the same value p has its starting node closest to R . Since the D-tree is deterministic, no two D-squares have the same value. Below we bind the number of D-squares in D-trees with the number of squares in ordinary trees. Recall that a *centroid* of a tree T is a node R such that each component of $T \setminus \{R\}$ contains at most $n/2$ nodes. It is a well-known fact that each tree has a centroid.

Lemma 4. *Assume that the number of D-squares in any D-tree of size n is $O(n^{4/3})$. Then the number of squares in any tree is also $O(n^{4/3})$.*

Proof. Let T be a tree of size n and let R be its centroid. Consider a D-tree $\mathcal{D} = (T_1, T_2, R)$ composed of two copies T_1 and T_2 of T , determined as in Lemma 3.

Let xx be a square in T anchored in R . Either this square or its reverse corresponds to a D-square in \mathcal{D} . Obviously $|\mathcal{D}| = O(n)$, therefore, by the hypothesis of the lemma, there are $O(n^{4/3})$ D-squares in this D-tree. Hence, the number of squares in T anchored in R is also $O(n^{4/3})$.

Now we need to count the squares in T that are not anchored in R . After removing the node R , the tree is partitioned into components T_1, \dots, T_k , such that $\sum_i |T_i| = n - 1$ and $|T_i| \leq n/2$. Hence, the number of squares in T can be written as:

$$\text{sq}(T) \leq O(|T|^{4/3}) + \sum_i \text{sq}(T_i).$$

A solution to this recurrence yields the upper bound $\text{sq}(n) = O(n^{4/3})$. \square

The proof of the assumption of the previous lemma is the core of this paper. In full generality it is provided in the last section. Here we limit ourselves to a very special type of squares. There is a useful connection between D-trees and combs, as expressed by the following observation, see Fig. 4.

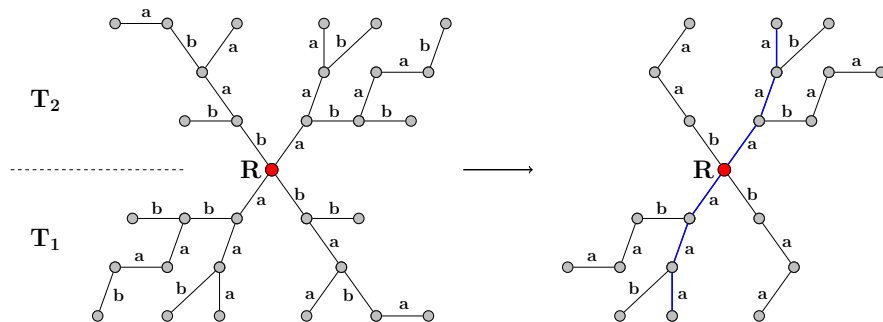


Fig. 4. Illustration of Observation 1.

Observation 1 Assume we have a D-tree labelled with letters a, b . Let us take only paths from a vertex in T_1 to R or from R to a vertex in T_2 which contain at most one b , with other edges labelled with a . Then the resulting labelled tree is a standard comb (with at most one additional branch attached to R).

Corollary 1. Assume binary alphabet $\{a, b\}$. The maximum number of special squares in any tree is $O(n^{4/3})$.

Proof. By Lemma 4, it suffices to consider a D-tree $\mathcal{D} = (T_1, T_2, R)$ and only special D-squares in \mathcal{D} . The special D-squares with both occurrences of b in T_2 are uniquely determined by their upper end and those with both occurrences in T_1 by their lower end. Hence the number of such D-squares is linear. By Lemma 2 and Observation 1, there are $O(n^{4/3})$ special D-squares which have one b in T_1 and one b in T_2 . \square

4 (p, q) -Representations of Substrings

In this section w denotes a word of length n . We start by recalling a few basic notions of word periodicity, see e.g. [5,6,11]. A *border* of w is defined as a prefix of w which is also a suffix of w . We say that a positive integer p is a *period* of $w = w_1w_2 \dots w_n$ if $w_i = w_{i+p}$ holds for all $1 \leq i \leq n - p$. A non-empty word w is called *periodic* if it has a period p such that $2p \leq |w|$. The *primitive root* of a word w , denoted $\text{root}(w)$, is the shortest word u such that $u^k = w$ for some positive integer k . We call a word w *primitive* if $\text{root}(w) = w$, otherwise it is called *non-primitive*. We recall several periodic properties of words [5,6,11].

Fact 1 A word w has a border of length b if and only if w has a period $|w| - b$.

Fact 2 (Periodicity Lemma) If a word of length n has two periods p and q , such that $p + q \leq n + \text{gcd}(p, q)$, then $\text{gcd}(p, q)$ is also a period of the word.

Fact 3 (Synchronizing Properties)

- (a) If $uv = vu$ then both words u, v are powers of the same primitive word.
- (b) Let $q \neq \varepsilon$ be a primitive word. Then q has exactly two occurrences in qq .
- (c) Let $p \neq \varepsilon, q \neq \varepsilon$ be such that pq is primitive. Then qp has exactly one occurrence in $pqqq$.

As a consequence of the synchronizing properties of primitive words, we obtain the following auxiliary fact that will be useful in the proof of the main result (Lemma 9).

Fact 4 Let p, p', q, q' be words such that: $q \neq \varepsilon$ and $q' \neq \varepsilon$, pq is primitive, $pq = p'q'$, and $qp = q'p'$. Then $p = p'$ and $q = q'$.

Proof. First assume $p = \varepsilon$. Then $q'p' = qp = pq = p'q'$. From Fact 3a, since $q' \neq \varepsilon$, we get $p' = \varepsilon$. This naturally implies that $q = q'$. Now assume that $p \neq \varepsilon$. We have $pqqq = p'q'p'q' = p'qpq'$ and from Fact 3c we know that there is only one occurrence of qp in $pqqq$. Thus $p = p'$ and $q = q'$. \square

Assume that w is periodic. There exists a unique representation of w : $w = (pq)^k p$ such that $k \geq 2$, $q \neq \varepsilon$ and pq is primitive. This representation is called a canonical representation of w . Here $|pq|$ is the shortest period of w . We say that w is of *periodic type* (p, q) .

Example 1. The word **abbabbab** has a canonical representation $(\mathbf{abb})^2 \mathbf{ab}$, with $p = \mathbf{ab}$ and $q = \mathbf{b}$. On the other hand, **bababa** has a representation $(\mathbf{ba})^3$ with $p = \varepsilon$ and $q = \mathbf{ba}$.

Fact 5 *Borders of w that are periodic belong to $O(\log n)$ periodic types. Additionally, w may have $O(\log n)$ borders which are not periodic.*

Proof. As for the first part of the lemma, let u, v be periodic borders of w such that $|u| < |v| \leq 1.5|u|$. We show that u and v are of the same periodic type.

Indeed, let $v = (pq)^k p$, where $d = |pq|$ is the shortest period of v , and $u = (p'q')^{k'} p'$ be the canonical representations of v and u . The border u is also a border of v . Due to Fact 1, both d and $|v| - |u|$ are periods of v . Moreover $d < \frac{1}{2}|v|$ (since $k \geq 2$) and $|v| - |u| \leq \frac{1}{3}|v|$ (since $|v| \leq 1.5|u|$). Hence, by the Periodicity Lemma, $|v| - |u|$ is a multiple of d . The word u is a prefix of v , hence $u = (pq)^\ell p$ for some $\ell < k$. Now, let us show that $\ell \geq 2$. Assume to the contrary that $\ell \leq 1$. Then:

$$3|p| + 2|q| \leq (k+1)|p| + k|q| = |v| \leq 1.5|u| \leq 3|p| + 1.5|q|.$$

This is clearly a contradiction, since $|q| > 0$. Hence $\ell \geq 2$. Now by the uniqueness of canonical representations we obtain $p = p'$, $q = q'$ and $k' = \ell$. This concludes that the borders u and v are of the same periodic type.

As for the second part, let u, v be non-periodic borders of w such that $|u| < |v|$. We show that $|v| > 2|u|$.

Assume to the contrary that $|u| < |v| \leq 2|u|$. As in the previous part of the proof, we see that $|v| - |u|$ is a period of v . However,

$$2(|v| - |u|) = |v| + |v| - 2|u| \leq |v| + 2|u| - 2|u| = |v|,$$

therefore v is periodic, a contradiction. □

A periodic border v of w is called *global* if its period is the period of the whole word w . Equivalently, v is global if v, w are of the same periodic type. If w is of periodic type (p, q) and its canonical representation is $w = (pq)^k p$, then all its global borders are $(pq)^{k'} p$ for $2 \leq k' \leq k$.

Definition 1. *Let p, q be such words that $q \neq \varepsilon$ and pq is primitive. The representation $w = p(qp)^\ell y(pq)^r p$ is called the (p, q) -representation of w if: (a) $\ell, r \geq 1$; (b) y has a prefix qp but not $(qp)^2$; (c) y has a suffix pq but not $(pq)^2$, see Fig. 5 and 6.*

Lemma 5. *Assume w has a non-global periodic border of periodic type (p, q) . Then w has a (p, q) -representation $w = p(qp)^\ell y(pq)^r p$. Moreover: (a) this (p, q) -representation is unique (i.e., ℓ, r and y are unique); (b) y is not a prefix of $(qp)^2$; (c) y is not a suffix of $(pq)^2$; (d) all borders of w of periodic type (p, q) are: $(pq)^{k'} p$ for $2 \leq k' \leq \min(\ell, r) + 1$.*

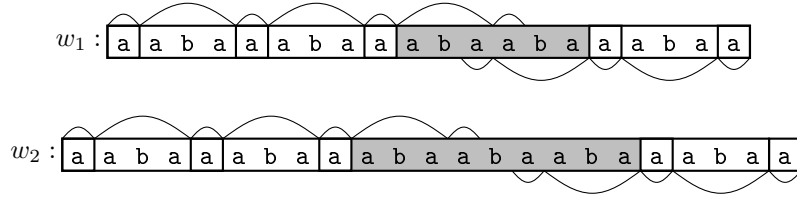


Fig. 5. The (p, q) -representations: $w_1 = \mathbf{a(abaa)^2abaaba(aaba)^1a}$ and $w_2 = \mathbf{a(abaa)^2abaabaaba(aaba)^1a}$. In both cases $p = \mathbf{a}$, $q = \mathbf{aba}$ and y is marked grey.

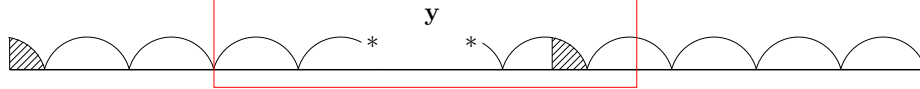


Fig. 6. A schematic view of a (p, q) -representation. The $*$ -symbols correspond to the first mismatch for the continuation of the period qp from the left side and the period pq from the right side.

Proof. Let $u = (pq)^k p$ be the longest border of w of type (p, q) . Clearly p is a prefix of w and $|w| > |u| > 2|p|$, so let us write $w = pzp$. Now, let $(qp)^{\ell+1}$ be the maximal power of qp that is a prefix of z . We have for some z' : $w = p(qp)^\ell z'p$. Let $(pq)^{r+1}$ be the maximal power of pq that is a suffix of z' . Now we can write $w = p(qp)^\ell y(pq)^r p$. Let us prove that this representation satisfies the required conditions. We get the following easily:

- z has a prefix qp and a suffix pq .
- z' has a prefix qp but not $(qp)^2$, and a suffix pq
- y has a prefix qp but not $(qp)^2$, and a suffix pq but not $(pq)^2$.

Let us now show that y is not a prefix of $(qp)^2$. Assume to the contrary. Recall that pq is a suffix of y . Thus we get an occurrence of pq in $qpqp$. If $p \neq \varepsilon$, from Fact 3 we get that $y = qpq$. But then, w would be of type (p, q) . Therefore $p = \varepsilon$. From Fact 3 we conclude that $y = q$ and w is a power of q . This is, however, impossible since w has a non-global periodic border $(pq)^k p = q^k$. This contradiction implies that y is not a prefix of $(qp)^2$.

A symmetric argument proves that y is not a suffix of $(pq)^2$. Since none of $(qp)^2$ and y is a prefix of the other, $p(qp)^{\ell+2}$ is not a prefix of w . Similarly $(pq)^{r+2}p$ is not a suffix of w . Thus $u = (pq)^{\min(\ell, r)+1}$. In particular $\ell, r \geq 1$. Clearly $(pq)^{k'} p$ for $2 \leq k' \leq \min(\ell, r) + 1$ are the only periodic borders of w of the type (p, q) .

Now it remains to show the uniqueness of the representation. Assume there was another representation $w = p(qp)^{\ell'} y' (pq)^{r'} p$. Since y' has a prefix qp but not $(qp)^2$, $\ell' + 1$ is the largest m such that $p(qp)^m$ is a prefix of w , that is $\ell' = \ell$. Similarly $r' = r$ and finally $y' = y$. \square

A periodic border is called *maximal* if it is the longest border of its periodic type. By Fact 5, w has $O(\log n)$ maximal borders.

We call a border *regular* if it is periodic and is neither global nor maximal.

5 General Combs and General Upper Bound

Due to Lemma 4, in this section we are only dealing with D-squares in a deterministic double tree $\mathcal{D} = (T_1, T_2, R)$ of size n .

For a node $v \in T_1$ we define the set $SQ(v)$ of all D-squares which start in v . Each D-square in $SQ(v)$ of value xx induces a period $|x|$ of $val(v, R)$, and thus corresponds to a border u of $val(v, R)$. This D-square is called regular if u is a regular border of $val(v, R)$. The periodic type of a D-square is defined as the periodic type of the underlying border u .

The following lemma lets us concentrate only on the regular D-squares.

Lemma 6. *At most $O(n \log n)$ D-squares in \mathcal{D} are not regular.*

Proof. We show that in $SQ(v)$ at most $O(\log n)$ D-squares are not regular. Each D-square in $SQ(v)$ corresponds to a different border of $val(v, R)$. The borders corresponding to non-regular D-squares are non-periodic, global or maximal; we extend these terms to D-squares as in the case of regular D-squares and borders. We have the following claim.

Claim. In $SQ(v)$ at most one D-square is global.

Proof. Let xx and $x'x'$ be values of two global D-squares starting in v . Assume $|x| < |x'|$. Let $w = val(v, R) = (pq)^k p$. The global D-squares are of the form $(pq)^{k'}$. Since a global border is periodic of periodic type (p, q) , we have $1 \leq k' \leq k - 2$. Let $x = (pq)^\ell$ and $x' = (pq)^{\ell'}$, $\ell < \ell'$.

Let u be an ancestor of v the defined by $val(u, R) = (pq)^{k-1} p$. A path starting in u and going to the upper end of $x'x'$ has the value $(pq)^{2\ell'-1}$, which has a prefix $(pq)^{2\ell} = xx$. We have $\ell < k - 1$, so this occurrence has a centre in the lower part of the D-tree. Hence, it is a candidate for a D-square of the value xx . This concludes that the original path of value xx starting in v could not be a D-square, which is a contradiction. \square

As we noticed in Section 4, only $O(\log n)$ borders of $val(v, R)$ can be non-periodic or maximal. Hence only $O(\log n)$ D-squares starting in v can correspond to a non-regular border. Thus there can be only $O(n \log n)$ D-squares which are not regular. \square

We introduce an important notion of a general comb. Before we give a formal definition, we provide a few sentences of intuition behind it. Assume $val(v, u)$ is a regular D-square of type (p, q) . By Lemma 5 we have the representation $val(v, R) = p(qp)^\ell y(pq)^r p$. All D-squares of type (p, q) starting in v correspond to the same representation. Those squares induce a particular structure of paths labelled with p, q and y in the upper part of the D-tree \mathcal{D} . A similar structure is also present in the lower part.

Definition 2. *Let p, q, y satisfy the conditions of Lemma 5. A D-tree (T_1, T_2, R) is called a (p, q, y) -comb if*

- for each leaf $v \in T_1$, $val(v, R) = p(qp)^m y(pq)^k p$ for some integers k, m ,

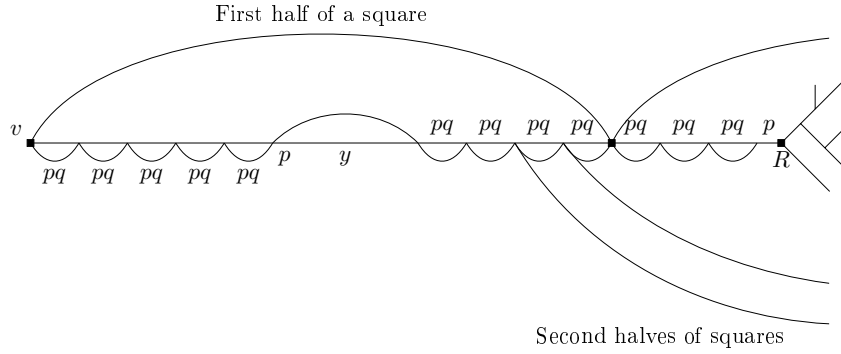


Fig. 7. Correspondence between squares in trees and borders.

– for each leaf $u \in T_2$, $val(R, u) = (qp)^m y (pq)^k$ for some integers k, m .

Let \mathcal{D} be a D -tree containing a regular D -square of periodic type (p, q) . Then by $Comb(\mathcal{D}, p, q, y)$ we denote the maximal subtree of \mathcal{D} that is a (p, q, y) -comb.

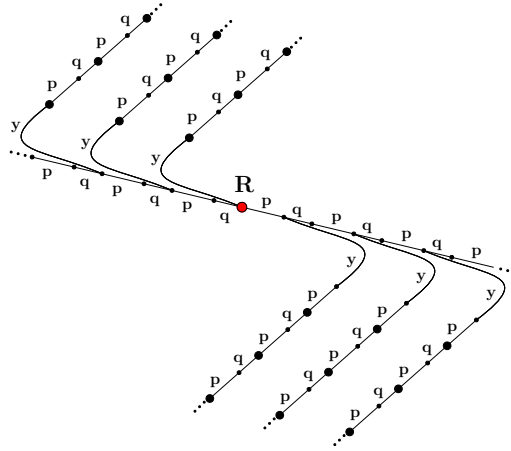


Fig. 8. A sample (p, q, y) -comb with the root R ; the main nodes are shown as larger circles; the bended edges are partially glued to the spine due to determinisation. All (p, q, y) -combs are subtrees of this infinite D -tree. For $p = \varepsilon$, $q = a$ and $y = aba$ we obtain a standard comb.

Note that the conditions of Definition 1 and Lemma 5 in particular imply that neither y is a prefix of $(qp)^2$ nor $(qp)^2$ a prefix of y . Similarly neither y is a suffix of $(pq)^2$ nor $(pq)^2$ a suffix of y . Hence, all combs have a regular structure, see Fig. 8. Each (p, q, y) -comb consists of a path labelled with $p(qp)^m$ (for an integer

m) and containing the root, which we call the *spine*, and the *branches* which are paths attached directly to the spine.

Some nodes of the combs are particularly important for the D-squares. These are nodes v of values $val(v, R) = p(qp)^k y(pq)^m p$ in the lower part, and nodes u of values $val(R, u) = (qp)^k y(pq)^m$ in the upper part (k, m are arbitrary nonnegative integers in both cases). Such nodes are called *main*. For a comb \mathcal{C} , by $Main(\mathcal{C})$ we denote the set of main nodes in \mathcal{C} , and by $\|\mathcal{C}\|$ we denote $|Main(\mathcal{C})|$. D-squares in \mathcal{C} with both endpoints in main nodes are said to be *induced* by the comb.

The following lemma confirms a strong relation between combs and regular D-squares.

Lemma 7. *Each regular D-square of type (p, q) in \mathcal{D} is induced by the corresponding comb $Comb(\mathcal{D}, p, q, y)$.*

Proof. Let $val(v, u)$ be a regular D-square in \mathcal{D} of type (p, q) . By Lemma 5, $val(v, R)$ has a following representation $val(v, R) = p(qp)^\ell y(pq)^r p$. The underlying border is regular, that is $(pq)^k p$ for some $2 \leq k \leq \min(\ell, r)$, hence the value of the D-square is $(p(qp)^\ell y(pq)^{r-k})^2$. Thus $val(R, u) = (qp)^{\ell-k} y(pq)^{r-k}$. Now it is clear that both v and u are main nodes of $Comb(\mathcal{D}, p, q, y)$. \square

The following result is a simple extension of the upper bound for standard combs.

Lemma 8. *A comb \mathcal{C} induces $O(\|\mathcal{C}\|^{4/3})$ D-squares.*

Proof. Let \mathcal{C} be a (p, q, y) -comb. We can construct a (ε, a, aba) -comb \mathcal{C}' of the same structure of branches and main nodes as \mathcal{C} . Clearly, $\|\mathcal{C}\| = \|\mathcal{C}'\|$ and the number of squares induced by both combs is the same. But now \mathcal{C}' is a standard comb.

For the comb \mathcal{C}' we have an upper bound $sq(\mathcal{C}') = O(\|\mathcal{C}'\|^{4/3})$ from Lemma 2. In order to obtain an $O(\|\mathcal{C}'\|^{4/3})$ bound for the number of squares *induced* by \mathcal{C}' , it suffices to restrict the proof of that lemma to special squares $(a^i b a^j)$ for $i \geq 2$ and $j \geq 1$. This way we obtain an upper bound of $O(\|\mathcal{C}'\|^{4/3})$ for the number of D-squares induced by \mathcal{C}' , consequently an $O(\|\mathcal{C}\|^{4/3})$ upper bound for an arbitrary comb \mathcal{C} . \square

Finally, we can prove the main lemma.

Lemma 9 (Key lemma). *A D-tree of size n contains $O(n^{4/3})$ regular D-squares.*

Proof. We show that combs in a D-tree are almost disjoint with regard to their main nodes. More precisely, due to combinatorial properties of words, any two different such combs can have at most two common main nodes in upper branches, and same for lower branches (Claim 2). Before that, we show that certain pairs of combs (with $|pq| = |p'q'|$) have none common main nodes at all (Claim 1).

Claim 1 *If $\mathcal{C} = Comb(\mathcal{D}, p, q, y)$ and $\mathcal{C}' = Comb(\mathcal{D}, p', q', y')$ are different combs satisfying $|pq| = |p'q'|$, then $Main(\mathcal{C}) \cap Main(\mathcal{C}') = \emptyset$.*

Proof. Assume u is a common main node of the two combs. It can lie either in the upper part or in the lower part of \mathcal{D} . First, let us consider the first case. Let $w = \text{val}(R, u)$. Since qp and $q'p'$ is a prefix of w and pq and $p'q'$ is a suffix of w , we get that $qp = q'p'$ and $pq = p'q'$. By Fact 4, $p = p'$ and $q = q'$. We now know that $w = (qp)^\ell y(pq)^r = (qp)^{\ell'} y'(pq)^{r'}$. Assume $\ell \neq \ell'$, without the loss of generality $\ell < \ell'$. Since y' has a prefix qp , $y(pq)^r$ has a prefix $(qp)^2$. This is impossible by the definition of a comb. By a similar argument, $r = r'$. Hence $y = y'$, so \mathcal{C} and \mathcal{C}' cannot be different combs.

Now, consider a common main node u in the lower part. Let $w = \text{val}(u, R)$. As previously we easily obtain that $p = p'$ and $q = q'$. This time we have $w = p(qp)^\ell y(pq)^r p = p(qp)^{\ell'} y'(pq)^{r'} p$. Exactly in the same way as before, we get $\ell = \ell'$, $r = r'$ and conclude that $y = y'$. \square

Claim 2 Let $\mathcal{C} = \text{Comb}(\mathcal{D}, p, q, y)$ and $\mathcal{C}' = \text{Comb}(\mathcal{D}, p', q', y')$. Then either $\mathcal{C} = \mathcal{C}'$ or $|\text{Main}(\mathcal{C}) \cap \text{Main}(\mathcal{C}')| \leq 4$.

Proof. By Claim 1, it suffices to show that if \mathcal{C} and \mathcal{C}' have at least 5 common main nodes, then $|pq| = |p'q'|$. First we show that no two common main nodes may lie on a single branch (in the upper or in the lower tree). Assume we have such two nodes u and u' and u is the lower among them. Then $\text{val}(u, u')$ is a power of both pq and $p'q'$. But pq and $p'q'$ are primitive, so $|pq| = |\text{root}(\text{val}(u, u'))| = |p'q'|$, which by claim concludes the proof of this case.

Now we show that no three common main nodes may lie in the upper tree. Assume that u, u', u'' are such nodes. Since no two of them can lie on the same branch, they are aligned as in Fig. 9 (up to a permutation of u, u', u'').

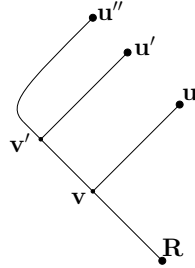


Fig. 9. Main nodes on three different branches of a D-tree.

Note that nodes v and v' need to be branching nodes of both combs. Obviously, $|pq| = |\text{root}(\text{val}(v, v'))| = |p'q'|$. This again concludes the proof of the current case.

Finally, it remains to show that no three common main nodes may lie in the lower tree. The proof is exactly the same as in the previous case. \square

Now let us divide combs into small combs, for which $\|\mathcal{C}\| \leq n^{0.6}$, and the remaining big combs. Due to the following claim, we can restrict the further analysis to big combs.

Claim 3 *The number of regular D-squares in \mathcal{D} induced by small combs is $o(n^{4/3})$.*

Proof. Consider a node v in the lower part of the D-tree \mathcal{D} . Assume $SQ(v)$ contains $s > 0$ regular D-squares of type (p, q) , and $val(v, R) = w = p(qp)^\ell y(pq)^r p$ is a corresponding representation. Let $\mathcal{C} = Comb(\mathcal{D}, p, q, y)$. We will show that $|Main(\mathcal{C})| = \Omega(s^2)$.

Indeed, let $x_1 x_1, \dots, x_s x_s$ be those s regular D-squares ordered by increasing lengths. As in the proof of Lemma 7 the values of these D-squares are of regular form. Namely, we have $x_i = p(qp)^\ell y(pq)^{k_i}$ for some $\max(0, r - \ell) \leq k_1 < \dots < k_s < r$. Let u_1, \dots, u_s be the other endpoints of these D-squares. We have $val(R, u_i) = (qp)^{\ell-r+k_i} y(pq)^{k_i}$. The nodes in the upper tree of \mathcal{C} corresponding to paths of the form $(qp)^{\ell-r+k_i} y(pq)^k$ for $0 \leq k \leq k_i$ are all distinct main nodes, hence $|Main(\mathcal{C})| \geq ((k_1 + 1) + (k_2 + 1) + \dots + (k_s + 1)) \geq (1 + 2 + \dots + s) = \Omega(s^2)$.

As a consequence, we get that $O(n^{0.3})$ D-squares from $SQ(v)$ can be induced by a single small comb. Moreover, by Fact 5, regular squares starting in v are induced by $O(\log n)$ combs. Consequently, the number of elements of $SQ(v)$ that are induced by small combs is $O(n^{0.3}) \cdot O(\log n) = o(n^{1/3})$. In total, small combs induce $o(n^{4/3})$ squares. \square

Let $\mathcal{C}_1, \dots, \mathcal{C}_k$ denote all big combs of \mathcal{D} . As a consequence of Claim 2, the total size of all these combs, measured in the number of main nodes, turns out to be linear in terms of n .

Claim 4 *For any D-tree of n nodes, $\sum_{i=1}^k \|\mathcal{C}_i\| = O(n)$.*

Proof. We will show the following inequality:

$$\sum_{i=1}^k \|\mathcal{C}_i\| \leq n + 2(k-1)(k-2). \quad (1)$$

From this inequality, by $\|\mathcal{C}_i\| \geq n^{0.6}$, we get

$$k \cdot n^{0.6} \leq n + 2(k-1)(k-2).$$

Comparing asymptotics of both sides of the inequality, we conclude that for *almost all* values of n (that is, all values excluding only a finite number) $k < n^{0.5}$. For such values of k the right side of the inequality (1) is $O(n)$, which will conclude the proof of the claim provided that we show that inequality.

As for the proof of (1), using Claim 2 we obtain that:

$$\begin{aligned} \left| \bigcup_{i=1}^k Main(\mathcal{C}_i) \right| &= \left| \bigcup_{i=1}^k \left(Main(\mathcal{C}_i) \setminus \bigcup_{j=1}^{i-1} Main(\mathcal{C}_j) \right) \right| \\ &= \sum_{i=1}^k \left| Main(\mathcal{C}_i) \setminus \bigcup_{j=1}^{i-1} Main(\mathcal{C}_j) \right| \\ &\geq \sum_{i=1}^k (\|\mathcal{C}_i\| - 4 \cdot (i-1)) = \sum_{i=1}^k \|\mathcal{C}_i\| - 2(k-1)(k-2). \end{aligned}$$

Consequently:

$$\sum_{i=1}^k \|\mathcal{C}_i\| - 2(k-1)(k-2) \leq \left| \bigcup_{i=1}^k \text{Main}(\mathcal{C}_i) \right| \leq n$$

which is equivalent to the inequality (1). \square

Let \mathcal{D} be a D-tree of size n . Due to Lemma 7, each regular D-square in \mathcal{D} is induced by a comb in \mathcal{D} . By Claim 3, there are $o(n^{4/3})$ such D-squares induced by small combs. Finally, by Lemma 8 and Claim 4, the number of regular D-squares induced by big combs $\mathcal{C}_1, \dots, \mathcal{C}_k$ of \mathcal{D} is bounded by:

$$\sum_{i=1}^k O\left(\|\mathcal{C}_i\|^{4/3}\right) = O\left(\sum_{i=1}^k \|\mathcal{C}_i\|\right)^{4/3} = O(n^{4/3}).$$

This completes the proof of the key lemma. \square

As a corollary of the key lemma, by Lemma 4 and 6 we obtain the desired upper bound.

Theorem 2. *The number of squares in a tree with n nodes is $O(n^{4/3})$.*

References

1. Noga Alon and Jaroslaw Grytczuk. Breaking the rhythm on graphs. *Discrete Mathematics*, 308(8):1375–1380, 2008.
2. Noga Alon, Jaroslaw Grytczuk, Mariusz Haluszczak, and Oliver Riordan. Non-repetitive colorings of graphs. *Random Struct. Algorithms*, 21(3-4):336–346, 2002.
3. Francine Blanchet-Sadri, Robert Mercas, and Geoffrey Scott. Counting distinct squares in partial words. *Acta Cybern.*, 19(2):465–477, 2009.
4. Bostjan Bresar, Jaroslaw Grytczuk, Sandi Klavzar, Staszek Niwczyk, and Iztok Peterin. Nonrepetitive colorings of trees. *Discrete Mathematics*, 307(2):163–172, 2007.
5. Maxime Crochemore, Christophe Hancart, and Thierry Lecroq. *Algorithms on Strings*. Cambridge University Press, 2007.
6. Maxime Crochemore and Wojciech Rytter. *Jewels of Stringology*. World Scientific, 2003.
7. A. S. Fraenkel and J. Simpson. How many squares can a string contain? *J. of Combinatorial Theory Series A*, 82:112–120, 1998.
8. Jaroslaw Grytczuk, Jakub Przybylo, and Xuding Zhu. Nonrepetitive list colourings of paths. *Random Struct. Algorithms*, 38(1-2):162–173, 2011.
9. Lucian Ilie. A simple proof that a word of length n has at most $2n$ distinct squares. *J. Comb. Theory, Ser. A*, 112(1):163–164, 2005.
10. Lucian Ilie. A note on the number of squares in a word. *Theor. Comput. Sci.*, 380(3):373–376, 2007.
11. M. Lothaire. *Combinatorics on Words*. Addison-Wesley, Reading, MA., U.S.A., 1983.