

# The Maximum Number of Square Substrings in a Tree

M.Crochemore<sup>1</sup> C. Iliopoulos<sup>1</sup> T.Kociumaka<sup>2</sup>

M.Kubica<sup>2</sup> J.Radoszewski<sup>2</sup> W.Rytter<sup>2</sup> W.Tyczynski<sup>2</sup>

T.Walen<sup>2</sup>

<sup>1</sup>King's College, London, UK.

<sup>2</sup>Warsaw University, Poland

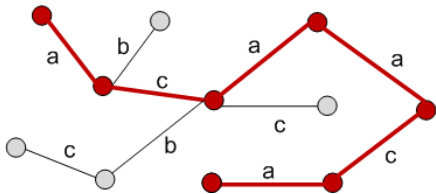
LSD 2012

## Abstract

We show that the maximum number of different square substrings in undirected labeled 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  $c, c' > 0$  we obtain:

$$c \cdot n^{1+\frac{1}{3}} \leq \text{sq}(n) \leq c' \cdot n^{1+\frac{1}{2}}.$$



There are 4 square substrings in this tree. The longest is *acaaca* and it corresponds to a path marked with a solid line.

In case of strings we can split the string into two halves and consider squares contained fully in one of these halves and squares overlapping the splitting **middle point**.

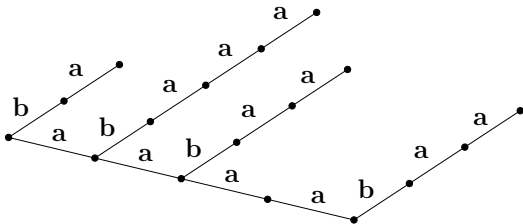
In case of tree the middle point corresponds to a **centroid**.

For a tree  $T$ , we define a *centroid* of  $T$  as a node  $R$  such that each component of  $T \setminus \{r\}$  contains at most  $n/2$  nodes.

## **Lemma**

*Each tree has a centroid that can be computed in linear time.*

A **comb** is a path of edges labeled with the letter *a* called the *backbone*, with a number of *branches* attached to different nodes in the backbone. Each such branch is a path starting with the letter *b*, followed by a number of *a*-labeled edges.



A comb containing 10 square substrings.

For  $m = k^2$  define:

$$Z_m = \{1, \dots, k\} \cup \{i \cdot k : 1 \leq i \leq k\}.$$

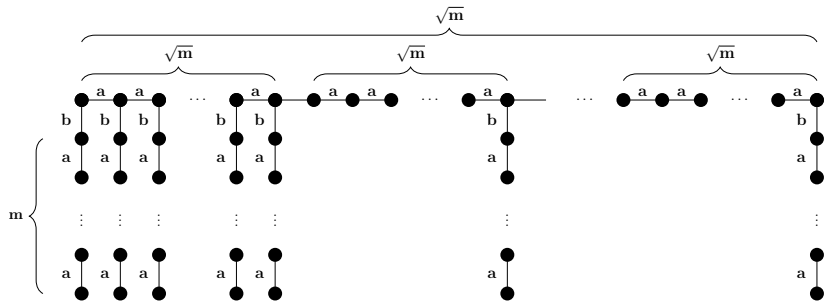
Each number  $0 < j < m$  can be written as  $p\sqrt{m} - q$ , where  $0 \leq p, q < \sqrt{m}$ .

### Lemma

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

**DEFINITION:** For  $m = k^2$  we define a comb  $T_m$  as follows:  $T_m$  consists of a backbone 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 backbone.

# The structure of a comb $T_m$



## Theorem

### [Lower Bound Theorem]

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

### Proof.

From the Lemma, for every  $1 < j < m$  there are two nodes  $u, v$  of degree 3 on the backbone, with  $distance(u, v) = j$ . Hence,  $T_m$  contains all squares of the form  $(a^i b a^{j-i})^2$  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  $\Omega(|T_m|^{4/3})$ . □



The squares of the form  $(a^i b a^i)^2$  correspond to paths ending in different branches, consequently we obtain the following observation.

### **Observation**

For each vertex  $v$  in a comb of size  $n$ , there are  $O(\sqrt{n})$  squares, which are values of paths starting at  $v$ . (This is a tight bound).

This gives simple proof of the fact:

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

We show much better upper bound.

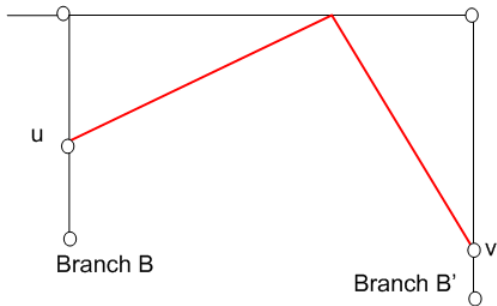
A comb contains only square substrings of the form  $(a^i)^2$  or  $(a^i b a^j)^2$ . The number of squares of the first type is  $O(n)$ . It is enough to bound the number of squares of the latter type, which we call *special* squares. Any occurrence of a special square starts and ends within two different branches of a comb. Note that there are at most  $n^{4/3}$  different special squares for which  $i < n^{2/3}$  and  $j < n^{2/3}$ .

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

A comb contains only square substrings of the form  $(a^i)^2$  or  $(a^i b a^j)^2$ . The number of squares of the first type is  $O(n)$ . It is enough to bound the number of squares of the latter type, which we call *special* squares. Any occurrence of a special square starts and ends within two different branches of a comb. Note that there are at most  $n^{4/3}$  different special squares for which  $i < n^{2/3}$  and  $j < n^{2/3}$ .

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

The square which starts at  $u$  can finish at other branch at most in one node  $v$ .



## Lemma

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

**PROOF** It is enough to show:

in any comb there are  $O(n^{4/3})$  (long) square substrings  $(a^i b a^j)^2$ ,

where  $i \geq n^{2/3}$  or  $j \geq n^{2/3}$

Consider a node  $u$  located in a branch  $B$  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  $B'$ .

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  $O(n^{4/3})$ .

## Lemma

*For each vertex  $v$  in a comb of size  $n$ , there are  $O(\sqrt{n})$  squares of the form  $(a^i b a^j)^2$ , which are values of paths starting at  $v$ .*

*(This is a tight bound).*

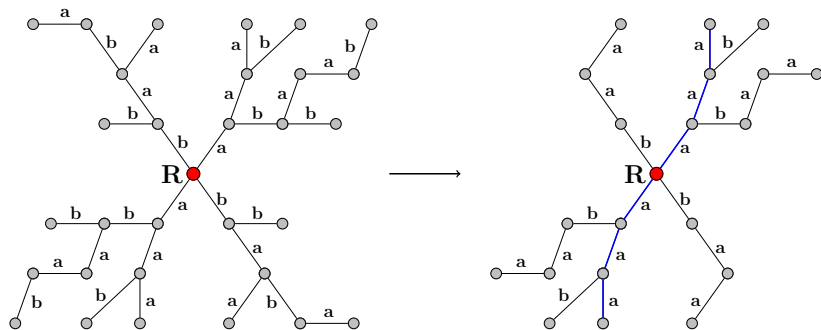
PROOF.

Assume  $v$  is not on the horizontal spine. They are going to different vertical branches, however at different heights. These branches are disjoint. If we have  $k$  such squares then we have  $k$  different disjoint parts of the comb and these parts are pairwise disjoint. Hence  $k \leq \sqrt{n}$ . The same argument applies to squares going from  $v$  to the left.

Define a *double tree*  $\mathcal{D} = (T_1, T_2, R)$  as a labeled 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  $|\mathcal{D}| = |T_1| + |T_2|$ . In the double tree we consider only squares which start in  $T_1$  and finish in  $T_2$ .

A directed rooted labeled tree is deterministic if the edges outgoing from the same vertex have different labels. 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*. Such determinism will prove quite essential later.

## D-tree and a derived comb



An example of a deterministic double tree (lower and upper trees), the anchor (root) is the red node. The squares of the form  $(a^*ba^*)^2$  with  $b$ 's in both trees from a comb.



## Lemma

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

## Proof.

Let us direct each tree  $T_i$  temporarily from  $R$  treated as a root. Assume we have a vertex  $v$  with edges  $(v, u)$ ,  $(v, w)$  going to its children and labeled with same letter  $a$ . Then we can *glue* the vertices  $u, w$ . We can do it going top-down from the root in a BFS traversal. □

We say that a path is anchored in  $R \in T$  iff  $R$  is an internal vertex of the path. A square is *anchored* in  $R$  iff it is a value of a path anchored in  $R$ .

## Lemma

**[Centroid Application Lemma]** *Assume we know that maximum number of squares in any D-tree of size  $n$  is  $O(n^\alpha)$ , where  $\alpha > 1$ . Then maximum number of squares in any tree is also  $O(n^\alpha)$ .*

PROOF.

Let  $T$  be a tree of size  $n$  and let  $R$  be its centroid. Consider a D-tree  $(T_1, T_2, R)$  composed of determinized two copies  $T_1$  and  $T_2$  of  $T$ ,

Now we need to count the squares in  $T$  that avoid the node  $R$ . After removing the node  $R$ , the tree is partitioned into small components  $T_1, \dots, T_k$ . Hence, the number of squares in  $T$  can be written as:

$$\text{sq}(T) = O(n^\alpha) + \sum_i \text{sq}(T_i), \text{ where}$$

$$\sum_i |T_i| = n - 1 \ \& \ |T_i| \leq n/2$$

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

There is a useful connection between D-trees and combs, as expressed by the following observation.

### Lemma

*Assume we have a D-tree labeled with letters  $a, b$ . Let us take only paths from each vertex in  $T_1, T_2$  to  $R$  which contain at most one  $b$ , with other edges labeled with  $a$ . Then the resulting labeled tree is a comb.*

We show only the upper bound for special squares. Assume binary alphabet  $\{a, b\}$ .

### Theorem

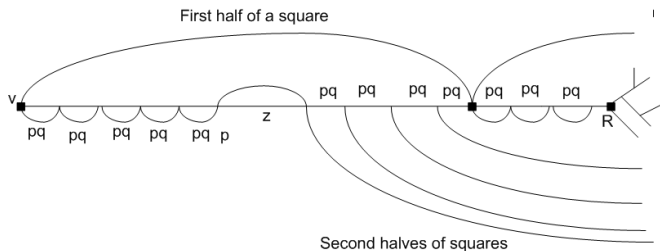
*The maximum number of special squares  $(a^*ba^*)^2$  in any tree is  $O(n^{4/3})$ .*

### Proof.

It suffices to consider D-trees  $\mathcal{D} = (T_1, T_2, R)$  and only special squares which have one letter  $b$  in  $T_1$  and one letter  $b$  in  $T_2$ . The number of other special squares going from  $T_1$  to  $T_2$  is linear.



Let  $\mathcal{D} = (T_1, T_2, R)$  be a D-tree. For a node  $v \in T_1$  we define the set  $SQ(v)$  of all squares corresponding to paths from  $T_1$  to  $T_2$  which start in  $v$  and their centers are between  $v$  and  $R$ . Each  $x \in SQ(v, R)$  induces a period  $|x|$  of  $val(v, R)$ , thus  $x$  admits a prefix that is a border of  $val(v, R)$ .



**Figure:** Correspondance between squares in trees and borders.

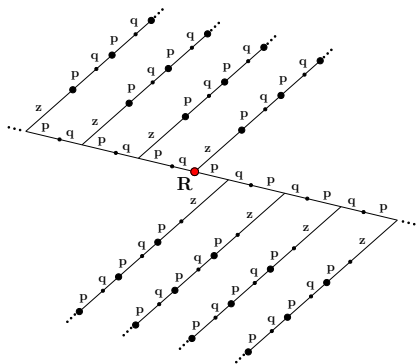
A  $(p, q, z)$ -comb is a D-tree that consists of a *spine* path labeled with  $(pq)^a p$  for some  $a \geq 0$  and *branches* of the form  $z(pq)^i$  for any non-negative integer  $i$  that can be attached after each  $pq$  on the spine. The root  $R$  of a comb divides the spine into non-empty parts  $(pq)^b p$  (lower part) and  $(pq)^{a-b}$  (upper part).

Hence, this occurrence of  $xx$  induces a particular structure in the D-tree corresponding to occurrences of segments  $p_iq_i$  and  $z_i$  ( $z_i = \varepsilon$  for  $i = 0$ ). This structure is called a  $(p_i, q_i, z_i)$ -comb in the D-tree rooted at  $R$ , denoted as  $COMB(p_i, q_i, z_i, \mathcal{D})$ .

Conversely, we say that  $xx$  is induced by this comb.

We assume that  $p, q, z$  are words (possibly empty) such that  $pq$  is non-empty and primitive,  $qp$  is *not* a prefix of  $z$  and  $pq$  is *not* a suffix of  $z$ . The above definition of a generalized comb is valid if  $z$  and  $qp$  start with a different letter. If  $z$  and  $qp$  have a common prefix  $x$ ,  $z = xz'$ ,  $qp = xw'$ , then the branches are glued to the spine: that is, they have the form  $z'(pq)^i$  and they can start between  $x$  and  $w'$  within each  $qp$ .





**Figure:** A  $(p, q, z)$ -comb with the root  $R$  corresponds to a standard comb by substituting  $pq \rightarrow a$ ,  $z \rightarrow b$ .

For each node  $v$  in the lower tree there are  $O(\log n)$  possible combs corresponding to squares originating from this node and having center in a lower tree. It can be shown that for a fixed  $v$  these combs are disjoint. We have shown before that the number of squares starting from a given vertex in a standard comb  $C$  is  $O(\sqrt{|C|})$ , this extends to generalized combs. Hence there are  $O(\sqrt{n})$  squares originating in a given, multiplying this by the number of vertices we get  $O(n^{3/2})$  squares in a D-tree. Due to the Centroid Application Lemma taking  $\alpha = 3/2$  we get.

### **Theorem**

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