Tight Bound for the Number of Distinct Palindromes in a Tree

Paweł Gawrychowski[1,⋆], Tomasz Kociumaka[1,⋆⋆], Wojciech Rytter[1,⋆⋆⋆], and Tomasz Waleń[1,†]

Faculty of Mathematics, Informatics and Mechanics,
University of Warsaw, Warsaw, Poland
[gawry,kociumaka,rytter,walen]@mimuw.edu.pl

Abstract. For an undirected tree with n edges labelled by single letters, we consider its substrings, which are labels of the simple paths between pairs of nodes. We prove that there are $\mathcal{O}(n^{1.5})$ different palindromic substrings. This solves an open problem of Brlek, Lafrenière and Provençal (DLT 2015), who gave a matching lower-bound construction. Hence, we settle the tight bound of $\Theta(n^{1.5})$ for the maximum palindromic complexity of trees. For standard strings, i.e., for paths, the palindromic complexity is $n + 1$.

1 Introduction

Regularities in words are extensively studied in combinatorics and text algorithms. One of the basic type of such structures are palindromes: words which are the same when read in both directions. The *palindromic complexity* of a word is the number of distinct palindromic substrings in the word. An elegant argument shows that the palindromic complexity of a word of length n does not exceed $n + 1$ [5], which is already attained by a unary word $\mathtt{a}^n$. Therefore the problem of palindromic complexity for words is completely settled, and the natural next step is to generalize it to trees.

In this paper we consider the palindromic complexity of undirected trees with edges labelled by single letters. We define substrings of such a tree as the labels of simple paths between arbitrary two nodes. Each label is the concatenation of the labels of all edges on the path. Fig. 1 illustrates palindromic substrings in a sample tree. Note that palindromes in a word of length n naturally correspond to palindromic substrings in a path of n edges.

⋆ Work done while the author held a post-doctoral position at Warsaw Center of Mathematics and Computer Science.

⋆⋆ Supported by Polish budget funds for science in 2013-2017 as a research project under the 'Diamond Grant' program.

⋆⋆⋆ This work was supported by the Polish National Science Center, grant no NCN2014/13/B/ST6/00770.

† Supported by the Polish Ministry of Science and Higher Education under the 'Iuventus Plus' program in 2015-2016 grant no 0392/IP3/2015/73.

The study of the palindromic complexity of trees was recently initiated by Brlek, Lafrenière and Provençal [3], who constructed a family of trees with n edges containing $\Theta(n^{1.5})$ distinct palindromic substrings. They conjectured that there are no trees with asymptotically larger palindromic complexity and proved this claim for a restricted case of trees in which the label of every path consists of up to 4 blocks (runs) of equal letters.

Our result. We show that the number of distinct palindromic substrings in a tree with n edges is $\mathcal{O}(n^{1.5})$. This bound is tight by the construction given in [3]; hence we completely settle the maximum palindromic complexity for trees.

Related work. Palindromic complexity of words was studied in various aspects. This includes algorithms determining the complexity [7], bounds on the average complexity [1] or generalizations to circular words [9]. Finite and infinite palindrome-rich words received particularly high attention; see e.g. [2,5,6]. This class contains, for example, all episturmian and thus all Sturmian words [5].

In the setting of labelled trees other kinds of regularities were also studied. It has been shown that a tree with n edges contains $\mathcal{O}(n^{4/3})$ distinct squares [4] and $\mathcal{O}(n \log n)$ distinct cubes [8]. The former bound is known to be tight. Interestingly, the lower bound construction resembles that for palindromes [3].

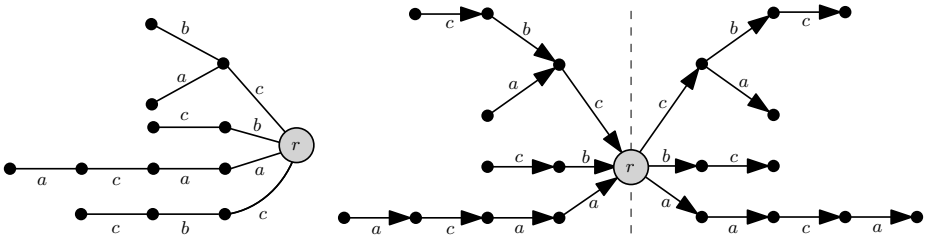

Fig. 1: To the left: an example undirected tree with 9 nontrivial palindromic substrings *bcb*, *bccb*, *aca*, *cbc*, *caac*, *cc*, *cbcbc*, *aa*, *acaaca*. To the right: deterministic double tree obtained after rooting the tree at r, merging both subtrees connected to r with edges labelled by c, and duplicating the resulting tree.

2 Preliminaries

A word w is a sequence of characters $w[1], w[2], \ldots, w[|w|] \in \Sigma$, often denoted $w[1..|w|]$. A substring of w is any word of the form $w[i..j]$, and if $i = 1$ ($j = |w|$) it is called a prefix (suffix). A period of w is any integer p, $1 \leq p \leq |w|$, such that $w[i] = w[i+p]$ for $i = 1, 2, \ldots, |w| - p$. The shortest period of w, denoted $\mathrm{per}(w)$, is the smallest such p. The following fact is a straightforward consequence of the periodicity lemma.

Fact 1. *Suppose a word v is a substring of a longer word u which has a period $p \leq \frac{1}{2}|v|$. Then* $\mathrm{per}(u) = \mathrm{per}(v)$.

A palindrome is a word w such that $w = w^R$, where w^R denotes the reverse of w. We have the following connection between periods and palindromes.

Fact 2. *Suppose a palindrome v is a suffix of a longer palindrome u. Then v is a prefix of u and thus $|u| - |v|$ is a period of u and of v.*

Define a *double tree* $\mathcal{D} = (T_\ell, T_r, r)$ as a labelled tree consisting of two trees T_ℓ and T_r sharing a common root r but otherwise disjoint. The edges of T_ℓ and T_r are directed to and from r, respectively. The size of $\mathcal{D}$ is defined as $|\mathcal{D}| = |T_\ell| + |T_r|$. For any $u, v \in \mathcal{D}$, we use $\mathrm{val}(u, v)$ to denote the sequence of the labels of edges on the path from u to v. A *substring* of $\mathcal{D}$ is a word $\mathrm{val}(u, v)$ such that $u \in T_\ell$ and $v \in T_r$. Also, let $\mathrm{d}(u, v) = |\mathrm{val}(u, v)|$ and $\mathrm{per}(u, v) = \mathrm{per}(\mathrm{val}(u, v))$.

We consider only *deterministic* double trees, meaning that all the edges outgoing from a node have distinct labels, and similarly all the edges incoming into a node have distinct labels. An example of such a double tree is shown in Fig. 1. Symmetry of palindromic substrings $\mathrm{val}(u, v)$, where $u \in T_\ell, v \in T_r$ gives a natural pairing of nodes on the path from u to v, where u is paired with v (and, if the path consists of an odd number of nodes, the central node is paired with itself). For any two paired nodes u', v' on such path, $\mathrm{val}(u', v')$ is a palindrome; if one of these two nodes is the root of the tree, we call the path from u' to v' the *central part* of the palindrome. Note that the central part is fully contained within T_ℓ or T_r. By symmetry of the counting problem (up to edge reversal in a double tree), we focus on palindromes admitting an occurrence whose central part lies in T_ℓ, or equivalently, occurring as $\mathrm{val}(u, v)$ with $\mathrm{d}(u, r) \geq \mathrm{d}(r, v)$.

3 Palindromes in Spine-Trees

A *spine-tree* is a deterministic double tree with a distinguished path, called *spine*, joining vertices $s_\ell \in T_\ell$ and $s_r \in T_r$. Additionally, we insist that this path cannot be extended preserving the period $p = \mathrm{per}(s_\ell, s_r)$. A palindromic substring is *induced* by such a spine-tree if its central part is a fragment of the spine of length at least p; see Fig. 2 for an example.

For a node u of the spine-tree let $s(u)$ denote the nearest node of the spine (if u is already on the spine, then $u = s(u)$). Since the spine-tree is deterministic, it satisfies the following property.

Fact 3. *For any induced palindrome* $\mathrm{val}(u, v)$*, the path* $\mathrm{val}(s(u), s(v))$ *is an inclusion-maximal fragment of* $\mathrm{val}(u, v)$ *admitting period p.*

Lemma 4. *There are up to $n\sqrt{n}$ distinct palindromic substrings induced by a spine-tree of size n.*

Proof. Define the *label* $L(u)$ for a node $u \in T_\ell$ as the prefix of $\mathrm{val}(u, s_r)$ of length $\mathrm{d}(u, s(u)) + p$. Similarly, the label $L(v)$ of a node $v \in T_r$ is the reversed suffix

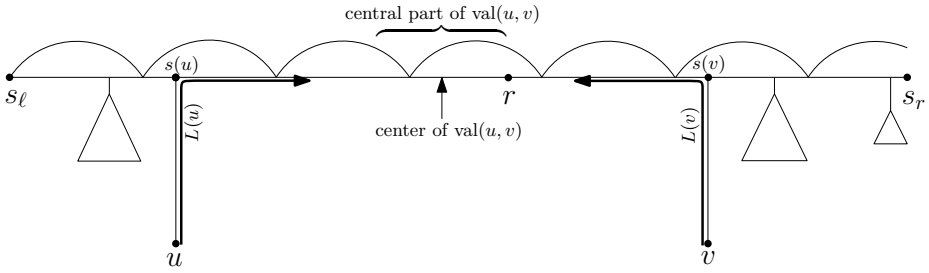

Fig. 2: A spine-tree, whose spine is the path from s_ℓ to s_r, with an induced palindrome $\mathrm{val}(u, v)$. Observe that $L(u) = L(v)$ is a prefix of the palindrome. Note that $d(s(u), r) \geq p$ but $d(r, s(v))$ might be smaller than p.

of $\mathrm{val}(s_\ell, v)$ of length $p + \mathrm{d}(s(v), v)$. We leave the label undefined if $\mathrm{val}(u, s_r)$ or $\mathrm{val}(s_\ell, v)$ is not sufficiently long, i.e., if $d(s(u), s_r) < p$ or $d(s_\ell, s(v)) < p$.

Consider a palindrome $\mathrm{val}(u, v)$ induced by the spine-tree. Fact 3 implies that $\mathrm{val}(s(u), s(v))$ is a maximal fragment of $\mathrm{val}(u, v)$ with period p. Since the central part of the palindrome is of length at least p and lies within this fragment, the fragment must be symmetric, i.e., we must have $\mathrm{d}(u, s(u)) = \mathrm{d}(s(v), v)$, and the labels of both u and v are defined. Consequently, $|L(u)| = |L(v)|$ and actually the labels $L(u)$ and $L(v)$ are equal. Hence, to bound the number of distinct palindromes, we group together nodes with the same labels. Let V_L be the set of vertices of $T_\ell \cup T_r$ with label L. We have the following claim.

Claim. For any label L, there are at most $\min(|V_L|^2, n)$ distinct palindromes with endpoints in V_L.

Proof. Consider all distinct induced palindromes $\mathrm{val}(u, v)$ such that $L(u) = L(v) = L$. A substring is uniquely determined by the endpoints of its occurrence, so $|V_L|^2$ is an upper bound on the number of these palindromes. We claim that every such palindrome is also uniquely determined by its length, which immediately gives the upper bound of n. Indeed $\mathrm{d}(u, s(u)) = \mathrm{d}(s(v), v) = |L| - p$ and $\mathrm{val}(s(u), s(v))$ has period p, so if the length is known, $\mathrm{val}(s(u), s(v))$ can be recovered from its prefix of length p, i.e., the suffix of L of length p. □

The sets V_L are disjoint, so by the above claim and using the inequality $\min(x, y) \leq \sqrt{xy}$ the number of distinct palindromes induced by the spine-tree is at most:

$$\sum_L \min(|V_L|^2, n) \leq \sum_L \sqrt{|V_L|^2 \cdot n} \leq \sqrt{n} \cdot \sum_L |V_L| \leq n^{1.5}. \qquad \square$$

4 Palindromes in General Deterministic Double Trees

Consider a node $u \in T_\ell$ and all distinct palindromes $P_1, \ldots, P_k$ with an occurrence starting at u. Observe that their central parts $C_1, \ldots, C_k$ have distinct

lengths: indeed, $|P_i| = 2\mathrm{d}(u,r) - |C_i|$ and $\mathrm{d}(u,r) \geq \frac{1}{2}|P_i|$, so $\mathrm{val}(u,r)$ and $|C_i|$ determines the whole palindrome P_i. Hence, we can order these palindromes so that $|C_1| > \ldots > |C_k|$, (i.e., $|P_1| < \ldots < |P_k|$).

Palindromes $P_{4\sqrt{n}+1}, \ldots, P_{k-2\sqrt{n}}$ are called *middle palindromes*. There are $\mathcal{O}(\sqrt{n})$ remaining palindromes for fixed u and $\mathcal{O}(n^{1.5})$ in total, so we can focus on counting middle palindromes. We start with the following characterization.

Lemma 5. *Consider middle palindromes $P_{4\sqrt{n}+1}, P_{4\sqrt{n}+2}, \ldots, P_{k-2\sqrt{n}}$ starting at node u. Central parts of these palindromes satisfy $|C_i| \geq 2\sqrt{n}$ and $\mathrm{per}(C_i) \leq \frac{1}{2}\sqrt{n}$. Moreover, for each P_i extending the central part C_i by $2\sqrt{n}$ characters in each direction preserves the shortest period.*

Proof. Since we excluded the $2\sqrt{n}$ palindromes with the shortest central parts, the middle palindromes clearly have central parts of length at least $2\sqrt{n}$.

First, let us prove that $\mathrm{per}(C_{2\sqrt{n}}) \leq \frac{1}{2}\sqrt{n}$. By Fact 2, $|C_j| - |C_{j+1}|$ is a period of C_j for $1 \leq j \leq 2\sqrt{n}$. Since $\sum_{j=1}^{2\sqrt{n}}(|C_j| - |C_{j+1}|) < |C_1| \leq n$, for some j we have $\mathrm{per}(C_j) \leq |C_j| - |C_{j+1}| \leq \frac{1}{2}\sqrt{n}$. Moreover, $C_{2\sqrt{n}}$ is a suffix of C_j, so the claim follows.

For $i > 4\sqrt{n}$, in particular if P_i is a middle palindrome, C_i is a suffix of $C_{2\sqrt{n}}$. Hence, Fact 1 implies that $\mathrm{per}(C_i) = \mathrm{per}(C_{2\sqrt{n}})$. Moreover, $|C_i| \leq |C_{2\sqrt{n}}| + 2\sqrt{n} - i < |C_{2\sqrt{n}}| - 2\sqrt{n}$, so extending C_i by $2\sqrt{n}$ characters to the left preserves the period. By symmetry of P_i, extension to the right also preserves the period. □

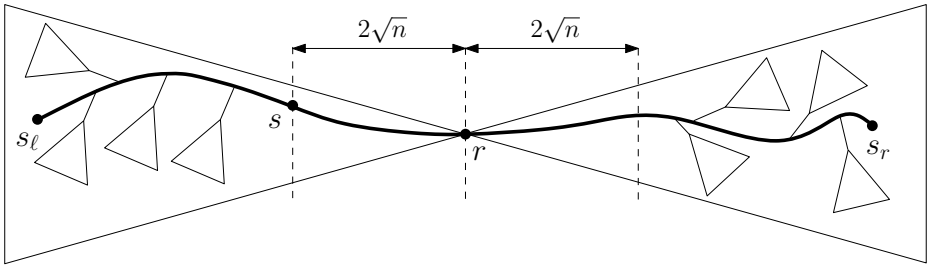

Fig. 3: A spine-tree constructed for a vertex s in a deterministic double tree. Note that we do not attach subtrees at distance less than $2\sqrt{n}$ from the root.

Let us choose any $s \in T_\ell$ such that $\mathrm{d}(s,r) = 2\sqrt{n}$ and $\mathrm{per}(s,r) \leq \frac{1}{2}\sqrt{n}$. Then, extend the period of $\mathrm{val}(s,r)$ to the left and to the right as far as possible, arriving at nodes s_ℓ and s_r, respectively. We create a spine-tree with spine corresponding to the path from s_ℓ to s_r as shown on Fig. 3. We attach to the spine all subtrees hanging off the original path at distance at least $2\sqrt{n}$ from the root. In other words, a vertex $u \in T_\ell$ which does not belong the spine is added to the spine-tree if $\mathrm{d}(s(u),r) \geq 2\sqrt{n}$ and a vertex $v \in T_r$ — if $\mathrm{d}(r,s(v)) \geq 2\sqrt{n}$. If $\mathrm{d}(r,s_r) < 2\sqrt{n}$ this leaves no subtrees hanging in T_r so we do not create any spine-tree for s.

Now consider a middle palindrome. By Lemma 5 its central part satisfies $|C| \geq 2\sqrt{n}$ and $\mathrm{per}(C) \leq \frac{1}{2}\sqrt{n}$. Moreover, by Fact 1 we have $\mathrm{per}(C) = \mathrm{per}(s, r)$ for the unique node $s \in T_\ell$ at distance $2\sqrt{n}$ from the root within C. Consequently, C lies on the spine of the spine-tree created for s and u belongs to a subtree attached to the spine. Additionally, since C can be extended by $2\sqrt{n}$ characters in each direction preserving the period, the other endpoint v must also belong to such a subtree in T_r (that is, we have $\mathrm{d}(r, s(v)) \geq 2\sqrt{n}$). Hence, any of the middle palindromic substrings is induced by some spine-tree.

The spine-trees are not disjoint, but nevertheless their total size is small.

Lemma 6. *The sizes $n_1, \ldots, n_k$ of the created spine-trees satisfy $\sum_i n_i \leq 2n$.*

Proof. We claim that at least $n_i - 2\sqrt{n}$ nodes of the i-th spine-tree are disjoint from all the other spine-trees. Let c_i be the node on the spine of the i-th spine-tree such that $\mathrm{d}(c_i, r) = \sqrt{n}$ and similarly let s_i satisfy $\mathrm{d}(s_i, r) = 2\sqrt{n}$. Recall that $\mathrm{per}(s_i, r) \leq \frac{1}{2}\sqrt{n}$. Thus, by Fact 1 $\mathrm{per}(s_i, r) = \mathrm{per}(c_i, r)$. Since the tree is deterministic, c_i uniquely determines s_i and hence the whole spine-tree. Thus, the nodes c_i are all distinct and so are their predecessors on the spines and all attached subtrees. A similar argument shows that all nodes d_i on the spine of the i-th spine-tree such that $\mathrm{d}(r, d_i) = \sqrt{n}$ are also all distinct. Therefore, we proved $\sum_i n_i - 2\sqrt{n} \leq n$. Each spine-tree has at least $4\sqrt{n}$ vertices on the spine, so this yields $n_i \geq 4\sqrt{n}$ and thus we obtain $\sum_i n_i \leq 2\sum_i (n_i - 2\sqrt{n}) \leq 2n$. □

By Lemma 4, the number of palindromes induced by the i-th spine-tree is at most $n_i^{1.5}$. Accounting the $\mathcal{O}(n^{1.5})$ palindromes which do not occur as middle palindromes, we have $\mathcal{O}(n^{1.5}) + \sum_i n_i^{1.5} \leq \mathcal{O}(n^{1.5}) + \sum_i n_i\sqrt{n} = \mathcal{O}(n^{1.5})$ palindromes in total.

Lemma 7. *Every deterministic double tree of size n has $\mathcal{O}(n^{1.5})$ distinct palindromic substrings.*

5 Main Result

To derive the final theorem, we follow the approach from [4]. We use the folklore fact that every tree T on n edges contains a *centroid* node r such that every component of $T \setminus \{r\}$ is of size at most $\frac{n}{2}$. We separately count palindromic substrings corresponding to the paths going through the centroid r and paths fully contained in a single component of $T \setminus \{r\}$. To bound the former, we root T at r directing all the edges so that they point towards the root, and then determinize the resulting tree by gluing together two children of the same node whenever their edges have the same label. Finally, we create a deterministic double tree by duplicating the tree and changing the directions of the edges in the second copy.

It is easy to see that for any simple path from u to v going through r in the original tree we can find $u' \in T_\ell$ and $v' \in T_r$ such that $\mathrm{val}(u, v) = \mathrm{val}(u', v')$. Hence, the number of distinct palindromic substrings corresponding to such

paths, by Lemma 7, is $\mathcal{O}(n^{1.5})$. Finally, we obtain the following recurrence for $\mathsf{pal}(n)$, the maximum number of palindromes in a tree with n edges:

$$\mathsf{pal}(n) = \mathcal{O}(n^{1.5}) + \max\left\{\sum_i \mathsf{pal}(n_i) : \forall_i\, n_i \leq \tfrac{n}{2} \;\wedge\; \sum_i n_i < n\right\}$$

which solves to $\mathsf{pal}(n) = \mathcal{O}(n^{1.5})$.

Theorem 8. *A tree with n edges contains $\mathcal{O}(n^{1.5})$ distinct palindromic substrings.*

References

1. Anisiu, M., Anisiu, V., Kása, Z.: Total palindrome complexity of finite words. Discrete Mathematics 310(1), 109–114 (2010)
2. Brlek, S., Hamel, S., Nivat, M., Reutenauer, C.: On the palindromic complexity of infinite words. Int. J. Found. Comput. Sci. 15(2), 293–306 (2004)
3. Brlek, S., Lafrenière, N., Provençal, X.: Palindromic complexity of tree-like graphs. In: Developments in Language Theorey. LNCS, Springer International Publishing (2015), http://arxiv.org/abs/1505.02695, to appear
4. Crochemore, M., Iliopoulos, C.S., Kociumaka, T., Kubica, M., Radoszewski, J., Rytter, W., Tyczyński, W., Waleń, T.: The maximum number of squares in a tree. In: Kärkkäinen, J., Stoye, J. (eds.) Combinatorial Pattern Matching. LNCS, vol. 7354, pp. 27–40. Springer Berlin Heidelberg (2012)
5. Droubay, X., Justin, J., Pirillo, G.: Episturmian words and some constructions of de Luca and Rauzy. Theor. Comput. Sci. 255(1-2), 539–553 (2001)
6. Glen, A., Justin, J., Widmer, S., Zamboni, L.Q.: Palindromic richness. Eur. J. Comb. 30(2), 510–531 (2009)
7. Groult, R., Prieur, É., Richomme, G.: Counting distinct palindromes in a word in linear time. Inf. Process. Lett. 110(20), 908–912 (2010)
8. Kociumaka, T., Radoszewski, J., Rytter, W., Waleń, T.: String powers in trees. In: Combinatorial Pattern Matching. LNCS, vol. 9133. Springer International Publishing (2015), to appear
9. Simpson, J.: Palindromes in circular words. Theor. Comput. Sci. 550, 66–78 (2014)