A Suffix Tree Or Not A Suffix Tree?

Tatiana Starikovskaya[1]

Higher School of Economics (Russia), tat.starikovskaya@gmail.com

Hjalte Wedel Vildhøj

Technical University of Denmark, DTU Compute, hwv@hwv.dk

Abstract

In this paper we study the structure of suffix trees. Given an unlabeled tree τ on n nodes and suffix links of its internal nodes, we ask the question "Is τ a suffix tree?", i.e., is there a string S whose suffix tree has the same topological structure as τ? We place no restrictions on S, in particular we do not require that S ends with a unique symbol. This corresponds to considering the more general definition of *implicit* or *extended* suffix trees. Such general suffix trees have many applications and are for example needed to allow efficient updates when suffix trees are built online. Deciding if τ is a suffix tree is not an easy task, because, with no restrictions on the final symbol, we cannot guess the length of a string that realizes τ from the number of leaves. And without an upper bound on the length of such a string, it is not even clear how to solve the problem by an exhaustive search. In this paper, we prove that τ is a suffix tree if and only if it is realized by a string S of length $n-1$, and we give a linear-time algorithm for inferring S when the first letter on each edge is known. This generalizes the work of I et al. [Discrete Appl. Math. 163, 2014].

Keywords: Reverse engineering; suffix trees; suffix links; suffix tour graphs.

1. Introduction

The suffix tree was introduced by Peter Weiner in 1973 [19] and remains one of the most popular and widely used text indexing data structures (see [1] and references therein). In static applications it is commonly assumed that suffix trees are built only for strings with a unique end symbol (often denoted \$), thus ensuring the useful one-to-one correspondance between leaves and suffixes. In this paper we view such suffix trees as a special case and refer to them as *\$-suffix trees*. Our focus is on suffix trees of *arbitrary strings*, which we simply call *suffix*

[☆] A preliminary version of this work has been presented at the 25th International Workshop on Combinatorial Algorithms in October 2014.

[1] Corresponding author. Partly supported by Dynasty foundation.

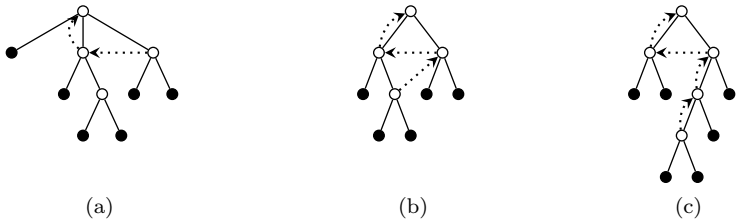


Figure 1: Three potential suffix trees. Internal nodes are black, and leaves are white. (a) is a \$-suffix tree, e.g. for `ababa$`. (b) is not a \$-suffix tree, but it is a suffix tree, e.g. for `abaabab`. (c) is not a suffix tree.

trees to emphasize that they are more general than \$-suffix trees[2]. Contrary to \$-suffix trees, the suffixes in a suffix tree can end in internal non-branching locations of the tree, called *implicit suffix nodes*.

Suffix trees for arbitrary strings are not only a nice generalization, but are required in many applications. For example in online algorithms that construct the suffix tree of a left-to-right streaming text (e.g., Ukkonen's algorithm [18]), it is necessary to maintain the implicit suffix nodes to allow efficient updates. Despite their essential role, the structure of suffix trees is still not well understood. For instance, it was only recently proved that each internal edge in a suffix tree can contain at most one implicit suffix node [4].

In this paper we prove some new properties of suffix trees and show how to decide whether suffix trees can have a particular structure. Structural properties of suffix trees are not only of theoretical interest, but are essential for analyzing the complexity and correctness of algorithms using suffix trees.

Given an unlabeled ordered rooted tree τ and suffix links of its internal nodes, the *suffix tree decision problem* is to decide if there exists a string S such that the suffix tree of S is isomorphic to τ. If such a string exists, we say that τ *is a suffix tree* and that S *realizes* τ. If τ can be realized by a string S having a unique end symbol \$, we additionally say that τ *is a \$-suffix tree*. See Fig. 1 for an example of a \$-suffix tree, a suffix tree, and a tree, which is not a suffix tree.

I et al. [15] recently considered the suffix tree decision problem and showed how to decide if τ is a \$-suffix tree in $O(n)$ time, assuming that the first letter on each edge of τ is also known. Deciding if τ is a suffix tree is much more involved, mainly because we can no longer infer the length of a string that realizes τ from the number of leaves. Without an upper bound on the length of such a string, it is not even clear how to solve the problem by an exhaustive search. In this paper, we give such an upper bound, show that it is tight, and give a linear time algorithm for deciding if τ is a suffix tree.

[2]In the literature the standard terminology is *suffix trees* for \$-suffix trees and *extended/implicit suffix trees* [3, 11] for suffix trees of strings not ending with \$.

Our Results. In Section 2, we start by settling the question of the sufficient length of a string that realizes τ.

Theorem 1. *An unlabeled tree τ on n nodes is a suffix tree if and only if it is realized by a string of length $n-1$.*

As far as we are aware, there were no previous upper bounds on the length of a shortest string realizing τ. The bound implies an exhaustive search algorithm for solving the suffix tree decision problem, even when the suffix links are not provided. In terms of n, this upper bound is tight, since e.g. stars on n nodes are realized only by strings of length at least $n-1$.

The main part of the paper is devoted to the suffix tree decision problem. We generalize the work of I et al. [15] and show in Section 4 how to decide if τ is a suffix tree.

Theorem 2. *Let τ be a tree with n nodes, annotated with suffix links of internal nodes and the first letter on each edge. There is an $O(n)$ time algorithm for deciding if τ is a suffix tree.*

In case τ is a suffix tree, the algorithm also outputs a string S that realizes τ. To obtain the result, we show several new properties of suffix trees, which may be of independent interest.

Related Work. The problem of revealing structural properties and exploiting them to recover a string realizing a data structure has received a lot of attention in the literature. Besides \$-suffix trees, the problem has been considered for border arrays [17, 7], parameterized border arrays [12, 13, 14], suffix arrays [2, 9, 16], KMP failure tables [8, 10], prefix tables [5], cover arrays [6], directed acyclic word graphs [2], and directed acyclic subsequence graphs [2].

2. Suffix Trees

In this section we prove Theorem 1 and some new properties of suffix trees, which we will need to prove Theorem 2. We start by briefly recapitulating the most important definitions.

The *suffix tree* of a string S is a compacted trie on suffixes of S [11]. Branching nodes and leaves of the tree are called *explicit nodes*, and positions on edges are called *implicit nodes*. The *label* of a node v is the string labelling the path from the root to v, and the length of this label is called the *string depth* of v. The *suffix link* of an internal explicit node v labeled by $a_1a_2\dots a_m$ is a pointer to the node u labeled by $a_2a_3\dots a_m$.

We use the notation $v \cdots\!\!\rightarrow u$ and extend the definition of suffix links to leaves and implicit nodes as well. We will refer to nodes that are labeled by suffixes of S as *suffix nodes*. All leaves of the suffix tree are suffix nodes, and unless S ends with a unique symbol \$, some implicit nodes and internal explicit nodes can be suffix nodes as well. Suffix links for suffix nodes form a path starting at the leaf labeled by S and ending at the root. Following [4], we call this path the *suffix chain*.

Lemma 1 ([4])**.** *The suffix chain of the suffix tree can be partitioned into the following consecutive segments: (1) Leaves; (2) Implicit suffix nodes on leaf edges; (3) Implicit suffix nodes on internal edges; and (4) Suffix nodes that coincide with internal explicit nodes. (See Fig. 2.)*

We define the parent $par(x)$ of a node x to be the deepest explicit node on the path from the root to x. The distance between a node and one of its ancestors is defined to be the difference between the string depths of these nodes.

Lemma 2. *If $x_1 \dashrightarrow x_2$ is a suffix link, then the distance from x_1 to $par(x_1)$ cannot be less than the distance from x_2 to $par(x_2)$.*

Proof. The string depth of x_2 is smaller than the string depth of x_1 by one. The end of the suffix link from $par(x_2)$ is an ancestor of x_2 and its string depth is smaller than the string depth of $par(x_1)$ by one again. Hence, the distance between x_1 and $par(x_1)$ is equal to the distance between x_2 and one of its ancestors, which is at least the distance between x_2 and $par(x_2)$. □

Lemma 3. *Let x be an implicit suffix node. The distance between x and $par(x)$ is not bigger than the length of any leaf edge.*

Proof. It follows from Lemma 2 that as the suffix chain $y_0 \dashrightarrow y_1 \dashrightarrow \ldots \dashrightarrow y_l = \text{root}$ is traversed, the distance from each node to its parent is non-increasing. Since the leaves are visited first, the distance between any implicit suffix node and its parent cannot exceed the length of a leaf edge. □

Lemma 4. *If τ is a suffix tree, then it can be realized by some string such that*

(1) The minimal length of a leaf edge of τ will be equal to one;

(2) Any edge of τ will contain at most one implicit suffix node at the distance one from its upper end.

Proof. Let S be a string realizing τ, and m be the minimal length of a leaf edge of τ. Consider a prefix S' of S obtained by deleting its last $(m-1)$ letters. Its suffix tree is exactly τ trimmed at height $m-1$. The minimal length of a leaf edge of this tree is one. Applying Lemma 3, we obtain that the distance between any implicit suffix node x of this tree and $par(x)$ is one, and, consequently, any edge contains at most one implicit suffix node. □

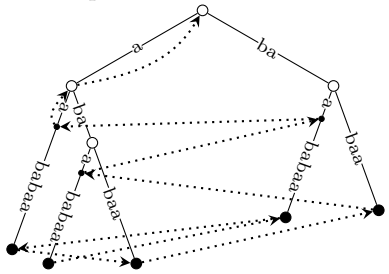


Figure 2: The suffix tree τ of a string $S = abaababaa$ with suffix nodes and the suffix chain. Suffix links of internal nodes are not shown.

Lemma 5. *If τ is realized by a string of length l, then it is also realized by strings of any length larger than l.*

Proof. Let $y_0 \dashrightarrow y_1 \dashrightarrow \ldots \dashrightarrow y_l = \text{root}$ be the suffix chain for a string S that realizes τ. Moreover let $letters(y_i)$ be the set of first letters immediately below node y_i. Then $letters(y_{i-1}) \subseteq letters(y_i)$, $i = 1, \ldots, l$. Let y_j be the first non-leaf node in the suffix chain (possibly the root). It follows that Sa also realizes τ, where a is any letter in $letters(y_j)$. □

Theorem 1. *An unlabeled tree τ on n nodes is a suffix tree if and only if it is realized by a string of length $n-1$.*

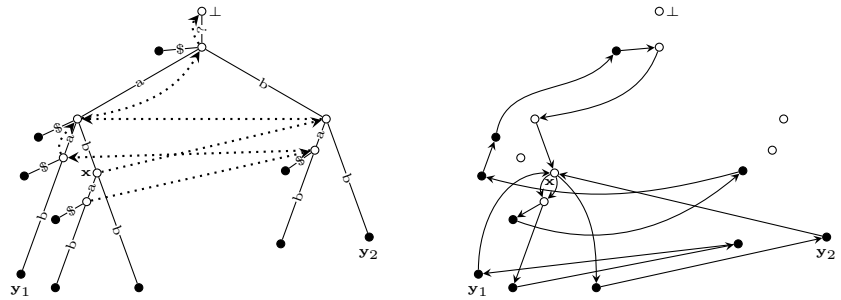
Proof. It suffices to show that if τ is a suffix tree then there is a string of length $n-1$ that realizes it. By Lemma 4, τ can be realized by a string S' so that the minimal length of a leaf edge is 1. Consider the last leaf ℓ visited by the suffix chain in the suffix tree of S'. By the property of S' the length of the edge $(par(\ell) \to \ell)$ is 1. Remember that a suffix link of an internal node always points to an internal node and that suffix links cannot form cycles. Moreover, upon transition by a suffix link the string depth decreases exactly by one. Hence if τ has I internal nodes then the string depth of the parent of ℓ is at most $I-1$ and the string depth of ℓ is at most I. Consequently, if L is the number of leaves in τ, the length of the suffix chain and thus the length of S' is at most $L + I - 1 = n - 1$, so by Lemma 5 there is a string of this length that realizes τ. □

3. Suffix Tour Graph

In their work [15] I et al. introduced a notion of *suffix tour graphs*. They showed that suffix tour graphs of \$-suffix trees must have a nice structure which ties together the suffix links of the internal nodes, the first letters on edges, and the order of leaves of τ — i.e., which leaf corresponds to the longest suffix, which leaf corresponds to the second longest suffix, and so on. Knowing this order and the first letters on edges outgoing from the root, it is easy to infer a string realizing τ. We study the structure of suffix tour graphs of suffix trees. We show a connection between suffix tour graphs of suffix trees and \$-suffix trees and use it to solve the suffix tree decision problem.

Let us first formalize the input to the problem. Consider a tree $\tau = (V, E)$ annotated with a set of suffix links $\sigma : V \to V$ between internal explicit nodes, and the first letter on each edge, given by a labelling function $\lambda : E \to \Sigma$ for some alphabet Σ. For ease of description, we will always augment τ with an auxiliary node $\perp$, the parent of the root. We add the suffix link (root $\dashrightarrow \perp$) to σ and label the edge ($\perp \to$ root) with a symbol "?", which matches any letter of the alphabet.

To construct the suffix tour of τ, we first compute values $\ell(x)$ and $d(x)$ for every explicit node x in τ. The value $\ell(x)$ is equal to the number of leaves y where $par(y) \dashrightarrow par(x)$ is a suffix link in σ, and $\lambda(par(y) \to y) = \lambda(par(x) \to x)$. Let L_x and V_x be the sets of leaves and nodes, respectively, of the subtree of



(a) (b)

Figure 3: (a) An input consisting of a tree, suffix links and the first letter on each edge extended with the auxiliary node $\perp$. (b) The corresponding suffix tour graph. The input (a) is the \$-suffix tree of a string `abaababaa`, which corresponds to an Euler tour of (b).

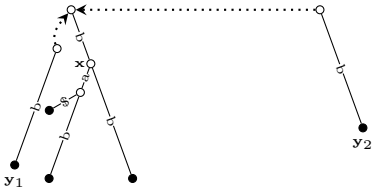


Figure 4: A fragment of the input tree above. For the node x we have $\ell(x) = 2$, because the parent of x is the endpoint of the suffix links for the parents of leaves y_1 and y_2. For all other nodes in the subtree of x the ℓ-values are equal to zero (no suffix links end at x or below). Since there are three leaves in the subtree of x, we have $d(x) = 1$. Consequently, in the suffix tour graph (see Fig. 3) x will have three incoming arcs: one from y_1, one from y_2, and one from the parent of x.

τ rooted at a node x. Note that L_x is a subset of V_x. We define $d(x) = |L_x| - \sum_{y \in V_x} \ell(y)$.

Definition 1. *The* suffix tour graph *of a tree $\tau = (V, E)$ is a directed graph $G = (V, E_G)$, where $E_G = \{(y \to x)^k \mid (y \to x) \in E, k = d(x)\} \cup \{(y \to x) \mid y$ is a leaf contributing to $\ell(x)\}$. If $k = d(x) < 0$, we define $(y \to x)^k$ to be $(x \to y)^{|k|}$. (See Fig. 3 and Fig. 4 for an example.)*

Below we show that suffix tour graphs of suffix trees are Eulerian graphs. The proof follows the lines of the proof of a similar claim for \$-suffix trees [15], but we revise it here for completeness and because of a different notation.

Lemma 6 ([15])**.** *The suffix tour graph G of a suffix tree τ is an Eulerian graph (possibly disconnected).*

Proof. To prove the lemma it suffices to show that for every node the number of incoming edges equals the number of outgoing edges. Consider an internal node x of τ. It has $\sum_{z \in children(x)} d(z)$ outgoing edges and $\ell(x) + d(x)$ incoming edges. But, $\ell(x) + d(x)$ equals

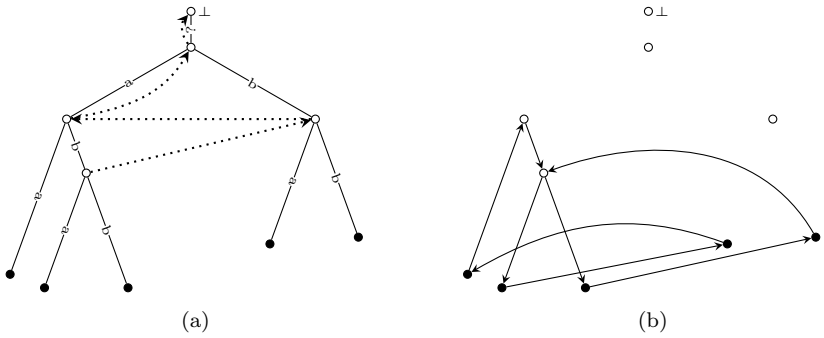


Figure 5: (a) An input consisting of a tree, suffix links and the first letter on each edge. The input is the suffix tree of `abaababaa` extended with the auxiliary node $\perp$. (b) The corresponding suffix tour graph.

$$|L_x| - \sum_{y \in V_x \setminus \{x\}} \ell(y) = \sum_{z \in children(x)} \big(|L_z| - \sum_{y \in V_z} \ell(y)\big) = \sum_{z \in children(x)} d(z)$$

Now consider a leaf y of τ. The outdegree of y is one, and the indegree is equal to $\ell(x) + d(x) = \ell(x) + 1 - \ell(x) = 1$. □

3.1. Suffix tour graph of a \$-suffix tree

The following proposition follows from the definition of a \$-suffix tree.

Proposition 1 ([15])**.** *If τ is a \$-suffix tree with a set of suffix links σ and first letters on edges defined by a labelling function λ, then*

(1) For every internal explicit node x in τ there exists a unique path $x = x_0 \dashrightarrow x_1 \dashrightarrow \ldots \dashrightarrow x_k =$ root such that $x_i \dashrightarrow x_{i+1}$ belongs to σ for all i;

(2) If y is the end of the suffix link for $par(x)$, there is a child z of y such that $\lambda(par(x) \to x) = \lambda(y \to z)$, and the end of the suffix link for x belongs to the subtree of τ rooted at y;

(3) For any node $x \in V$ the value $d(x) \geq 0$.

If all tree conditions hold, it can be shown that

Lemma 7 ([15])**.** *The tree τ is a \$-suffix tree iff its suffix tour graph G contains a cycle C which goes through the root and all leaves of τ. Moreover, a string realizing τ can be inferred from C in linear time.*

In more detail, the authors of [15] proved that the order of leaves in the cycle C corresponds to the order of suffixes. That is, the i^{th} leaf after the root corresponds to the i^{th} longest suffix. Thus, the string can be reconstructed in linear time: its i^{th} letter will be equal to the first letter on the edge in the path from the root to the i^{th} leaf. Note that the cycle and hence the string is not necessarily unique. See Fig. 3 for an example.

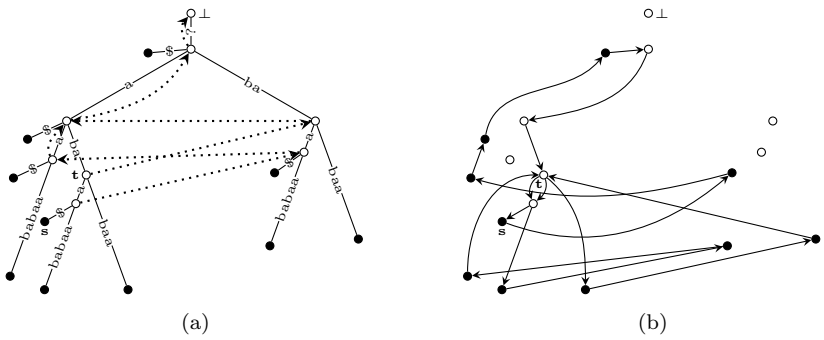


Figure 6: (a) The \$-suffix tree of `abaababaa`. (b) The corresponding suffix tour graph. Labels **s** and **t** mark the deepest \$-leaf and the twist node, respectively.

3.2. Suffix tour graph of a suffix tree

The high level idea of our solution is to try to augment the input tree so that the augmented tree is a \$-suffix tree. More precisely, we will try to augment the suffix tour graph of the tree to obtain a suffix tour graph of a \$-suffix tree. It will be essential to understand how the suffix tour graphs of suffix trees and \$-suffix trees are related.

Let ST and $ST_\$$ be the suffix tree and the \$-suffix tree of a string. We call a leaf of $ST_\$$ a \$-*leaf* if the edge ending at it is labeled by a single letter \$. Note that to obtain $ST_\$$ from ST we must add all \$-leaves, their parents, and suffix links between the consecutive parents to ST. We denote the deepest \$-leaf by s. (See Fig. 6.)

An internal node x of a suffix tour graph has $d(x)$ incoming arcs produced from edges and $\ell(x)$ incoming arcs produced from suffix links. All arcs outgoing from x are produced from edges, and there are $d(x) + \ell(x)$ of them since suffix tour graphs are Eulerian graphs. A leaf x of a suffix tour graph has $d(x)$ incoming arcs produced from edges, $\ell(x)$ incoming arcs produced from suffix links, and one outgoing arc produced from a suffix link. Below we describe what happens to the values $d(x)$ and $\ell(x)$, and to the outgoing arcs produced from suffix links. These two things define the changes to the suffix tour graph.

Lemma 8. *For the deepest \$-leaf s we have $\ell(s) = 0$ and $d(s) = 1$. The ℓ-values of other \$-leaves are equal to one, and their d-values are equal to zero. (See Fig. 6.)*

Proof. Suppose that $\ell(s) = 1$. Then there is a leaf y such that $par_\$(y) \cdots\!\!\rightarrow par_\$(s)$ is a suffix link in σ, and the first letter on the edge from $par_\$(y)$ to y is \$. That is, y is a \$-leaf and its string depth is bigger than the string depth of s, which is a contradiction. Hence, $\ell(s) = 0$ and therefore $d(s) = 1$. The parent of any other \$-leaf y will have an incoming suffix link from the parent of the previous \$-leaf and hence $\ell(y) = 1$ and $d(y) = 0$. □

The important consequence of Lemma 8 is that in the suffix tour graph of $ST_\$$ all the \$-leaves are connected by a path starting in the deepest \$-leaf and ending in the root. (See Fig. 6.)

Next, we consider nodes that are explicit in ST and $ST_\$$. If a node x is explicit in both trees, we denote its (explicit) parent in ST by $par(x)$ and in $ST_\$$ — by $par_\$(x)$. Below in this section we assume that each edge of ST contains at most one implicit suffix node at distance one from its parent.

Lemma 9. *Consider a node x of ST. If a leaf y contributes to $\ell(x)$ either in ST or $ST_\$$, and $par_\$(y)$ and $par_\$(x)$ are either both explicit or both implicit in ST, then y contributes to $\ell(x)$ in both trees.*

Proof. If $par_\$(y)$ and $par_\$(x)$ are explicit, the claim follows straightforwardly.

Consider now the case when $par_\$(y)$ and $par_\$(x)$ are implicit. Suppose first that y contributes to $\ell(x)$ in $ST_\$$. Then the labels of $par_\$(y)$ and $par_\$(x)$ are La and $L[2..]a$ for some string L and a letter a. Remember that distances between $par_\$(y)$ and $par(y)$ and between $par_\$(x)$ and $par(x)$ are equal to one. Therefore, labels of $par(y)$ and $par(x)$ are L and $L[2..]$, and the first letters on edges $par(x) \to x$ and $par(y) \to y$ are equal to a. Consequently, $par(y) \cdots\!\!\rightarrow par(x)$ is a suffix link, and y contributes to $\ell(x)$ in ST as well.

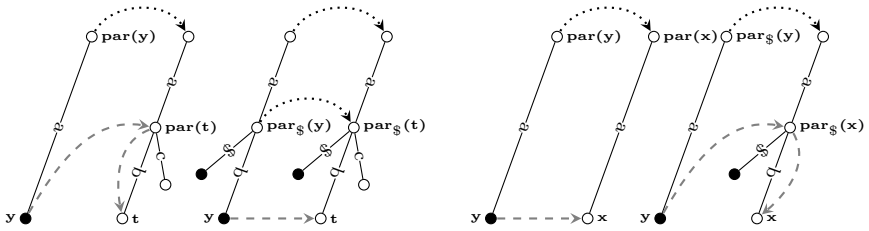
Now suppose that y contributes to $\ell(x)$ in ST. Then the labels of $par(y)$ and $par(x)$ are L and $L[2..]$, and the first letters on the edges $par(y) \to y$ and $par(x) \to x$ are equal to some letter a. This means that the labels of $par_\$(y)$ and $par_\$(x)$ are La and $L[2..]a$, and hence there is a suffix link from $par_\$(y)$ to $par_\$(x)$. Since y and x are not \$-leaves, y contributes to $\ell(x)$ in $ST_\$$. □

Before we define the deepest \$-leaf s. If the parent of s is implicit in ST, the changes between ST and $ST_\$$ are more involved. To describe them, we first need to define the *twist node*. Let p be the deepest explicit parent of any \$-leaf in ST. The node that precedes p in the suffix chain is thus an implicit node in ST, i.e., it has two children in $ST_\$$, one which is a \$-leaf and another node is y, which is either a leaf or an internal node. If y is a leaf, let t be the child of p such that y contributes to $\ell(t)$. We refer to t as the *twist* node. (See Fig. 6.)

Lemma 10. *Let x be a node of ST. Upon transition from ST to $ST_\$$, the ℓ-value of $x = t$ increases by one and the ℓ-value of its parent decreases by one. If $par_\$(x)$ is an implicit node of ST, then $\ell(x)$ decreases by $\ell(par_\$(x))$. Otherwise, $\ell(x)$ does not change.*

Proof. The value $\ell(x)$ can change when (1) A leaf y contributes to $\ell(x)$ in $ST_\$$, but not in ST; or (2) A leaf y contributes to $\ell(x)$ in ST, but not in $ST_\$$.

In the first case the nodes $par_\$(y)$ and $par_\$(x)$ cannot be both explicit or both implicit. Moreover, from the properties of suffix links we know that if $par_\$(y)$ is explicit in ST, then $par_\$(x)$ is explicit as well [11]. Consequently, $par_\$(y)$ is implicit in ST, and $par_\$(x)$ is explicit. Since $par_\$(x)$ is the first explicit suffix node and y is a leaf that contributes to $\ell(x)$, we have $x = t$, and $\ell(x) = \ell(t)$ in $ST_\$$ is bigger than $\ell(t)$ in ST by one (see Fig. 7a).



(a) (b)

Figure 7: Both figures show ST on the left and $ST_\$$ on the right. Arcs of the suffix tour graphs that change because of the twist node t (Fig. 7a) and because of an implicit parent (Fig. 7b) are the grey dashed arcs.

Consider one of the leaves y satisfying (2). In this case $par(y) \dashrightarrow par(x)$ is a suffix link, and the first letters on the edges $par(y) \to y$ and $par(x) \to x$ are equal. Since y does not contribute to $\ell(x)$ in $ST_\$$, exactly one of the nodes $par_\$(y)$ and $par_\$(x)$ must be implicit in ST. Hence, we have two subcases: (2a) $par_\$(y)$ is implicit in ST, and $par_\$(x)$ is explicit; (2b) $par_\$(y)$ is explicit in ST, and $par_\$(x)$ is implicit.

In the subcase (2*a*) the distance between $par(y)$ and $par_\$(y)$ is one. The end of the suffix link for $par_\$(y)$ must belong to the subtree rooted at x. From the other hand, the string distance from $par(x)$ to the end of the suffix link is one. This means that the end of the suffix link is x. Consequently, x is the parent of the twist node t, and the value $\ell(x) = \ell(par_\$(t))$ is smaller by one in $ST_\$$ (see Fig. 7a).

In the subcase (2*b*) the ℓ-value of x in ST is bigger than the ℓ-value of x in $ST_\$$ by $\ell(par_\$(x))$, as all leaves contributing to $par_\$(x)$ in $ST_\$$, e.g. y, switch to x in ST (see Fig. 7b). □

Lemma 11. *Let x be a node of ST. Upon transition from ST to $ST_\$$, the value $d(x)$ of a node x such that $par_\$(x)$ is implicit in ST increases by $\ell(par_\$(x))$. If x is the twist node t, its d-value decreases by one. Finally, the d-values of all ancestors of the deepest \$-leaf s increase by one.*

Proof. Remember that $d(x) = |L_x| - \sum_{y \in V_x} \ell(y)$. If $par_\$(x)$ is implicit in ST, $\ell(x)$ decreases by $\ell(par_\$(x))$, i.e. $d(x)$ increases by $\ell(par_\$(x))$. Note that d-values of ancestors of x are not affected since for them the decrease of $\ell(x)$ is compensated by the presence of $par_\$(x)$. The value $\ell(t)$ increases by one and results in decrease of $d(t)$ by one, but for other ancestors of t increase of $\ell(t)$ will be compensated by decrease of $\ell(par_\$(t))$.

The value $\ell(s) = 0$ and the ℓ-values of other \$-leaves are equal to one. Consequently, when we add the \$-leaves to ST, d-values of ancestors of s increase by one, and d-values of ancestors of other \$-leaves are not affected. □

Lemma 12. *Let $par_\$(x)$ be an implicit parent of a node $x \in ST$. Then $d(par_\$(x))$ in $ST_\$$ is equal to $d(x)$ in ST if the node $par_\$(x)$ is not an ancestor of s, and $d(x) + 1$ otherwise.*

Proof. First consider the case when $par_\$(x)$ is not an ancestor of s. Remember that the suffix tour graph is an Eulerian graph. The node $par_\$(x)$ has $\ell(par_\$(x))$ incoming arcs produced from suffix links and $d(x)$ outgoing arcs produced from edges. Hence it must have $d(x) - \ell(par_\$(x))$ incoming arcs produces from edges, and this is equal to $d(x)$ in ST. If $par_\$(x)$ is an ancestor of s, the d-value must be increased by one as in the previous lemma. □

Speaking in terms of suffix tour graphs, we make local changes when the node is the twist node t or when the parent of a node is implicit in ST, and add a cycle from the root to s (increase of d-values of ancestors of s) and back via all \$-leaves.

4. Decision Algorithm

We are now ready to show the main result of this paper: A linear-time algorithm that decides if a given tree is a suffix tree.

Theorem 2. *Let τ be a tree with n nodes, annotated with suffix links of internal nodes and the first letter on each edge. There is an $O(n)$ time algorithm for deciding if τ is a suffix tree.*

Proof. We replace the original problem with the following one: Can τ be augmented to become a \$-suffix tree? We assume that τ satisfies Observation 1(1) and Observation 1(2), which can be verified in linear time. We will not violate these conditions while augmenting τ.

Remember that Observation 1(1) guarantees that for each node there is a unique suffix link path going from it to the root. We start by computing lengths of suffix links paths for each node, which can be done in linear time. If τ is a suffix tree, the lengths are equal to the string depths of the nodes.

The deepest \$-leaf s can either hang from a node of τ, or from an implicit suffix node $par_\$(s)$ on an edge of τ. Furthermore, if τ is a suffix tree, then we can assume that it has all the properties described in Lemma 4. Consequently, if $par_\$(s)$ is implicit in τ, the distance from $par_\$(s)$ to the upper end of the edge is equal to one. That is, there are $O(n)$ possible locations of s. For each of the locations we consider a suffix link path starting at its parent. The suffix link paths form a tree which we refer to as the suffix link tree. The suffix link tree can be built in linear time: For explicit locations the paths already exist, and for implicit locations we can build the paths following the suffix link path from the upper end of the edge containing a location and exploiting the knowledge about lengths of internal edges. If we see a node encountered before, the algorithm stops.

If τ is a suffix tree, then it is possible to augment it so that its suffix tour graph will satisfy Observation 1(3) and Lemma 7. We remind that Observation 1(3) says that for any node x of the suffix tour graph $d(x) \geq 0$, and Lemma 7 says that the suffix tour graph contains a cycle going through the root and all leaves. We show that each of the conditions can be verified for all possible ways

to augment τ by a linear time traverse of τ or the suffix link tree. We start with Observation 1(3).

Lemma 13. *If τ can be augmented to become a \$-suffix tree, then for all nodes x we have $d(x) \geq -1$.*

Proof. The value $d(x)$ increases only when x is an ancestor of s or when $par_{\$}(x)$ is implicit in ST. In the first case it increases by one. Consider the second case. Remember that $d(par_{\$}(x))$ is equal to $d(x)$ or to $d(x)+1$ if it is an ancestor of s. Since in a \$-suffix tree all d-values are non-negative, we have $d(x) \geq -1$ for any node x. □

Step 1. We first compute all d-values and all ℓ-values. If $d(x) \leq -2$ for some node x of τ, then τ cannot be augmented to become a \$-suffix tree and hence it is not a suffix tree. From now on we assume that τ does not contain such nodes. All nodes x with $d(x) = -1$, except for at most one, must be ancestors of s. If there is a node with a negative d-value that is not an ancestor of s, then it must be the lower end of the edge containing $par_{\$}(s)$, and the d-value must become non-negative after we augment τ.

We find the deepest node x with $d(x) = -1$ by a linear time traverse of τ. All nodes with negative d-values must be its ancestors, which can be verified in linear time. If this is not the case, τ is not a suffix tree. Otherwise, the possible locations for the parent of s are descendants of x and the implicit location on the edge to x if $d(x)+\ell(x)$, the d-value of x after augmentation, is at least zero. We cross out all other locations.

Remark 1. *Given two nodes of τ we can determine if one is an ancestor of the other in constant time after a standard linear-time preprocessing of τ.*

Step 2. For each of the remaining locations we consider the suffix link path starting at its parent. If the implicit node q preceding the first explicit node p in the path belongs to a leaf edge then the twist node t is present in τ and will be a child of p. We cannot tell which child though, since we do not know the first letter on the leaf edge outgoing from q. However, we know that $d(t)$ decreases by 1 after augmentation, and hence $d(t)$ must be at least 0. Moreover, if $d(t) = 0$ the twist node t must be an ancestor of s to compensate for the decrease of $d(t)$ ($\star$).

For each of the locations of s we check if t exists, and if it does, we find the parent of t. This can be done in linear time by a traverse of the suffix link tree. We then select the locations of s such that the parent of t is an ancestor of s or it has at least one child satisfying ($\star$). This can be done in linear time as well.

Step 3. We assume that the suffix tour graph of τ is an Eulerian graph, otherwise τ is not a suffix tree by Lemma 7. This condition can be verified in linear time. When we augment τ, we add a cycle C from the root to the deepest \$-leaf s and back via \$-leaves. The resulting graph will be an Eulerian graph as well, and one of its connected components (cycles) must contain the root and all leaves of τ.

We divide C into three segments: the path from from the root to the parent $par(x)$ of the deepest node x with $d(x) = -1$, the path from $par(x)$ to s, and the path from s to the root. We start by adding the first segment to the suffix tour graph. This segment is present in the cycle C for any choice of s, and it might actually increase the number of connected components in the graph. (Remember that if C contains an edge $x \to y$ and the graph contains an edge $y \to x$, then the edges eliminate each other.)

The second segment cannot eliminate any edges of the graph, and if it touches a connected component then all its nodes are added to the component containing the root of τ. Since the third segment contains the \$-leaves only, the second segment must go through all connected components that contain leaves of τ. We paint nodes of each of the components into some color. And then we perform a depth-first traverse of τ maintaining a counter for each color and the total number of distinct colors on the path from the root to the current node. When a color counter becomes equal to zero, we decrease the total number of colors by one, and when a color counter becomes positive, we increase the total number of colors by one. If a possible location of s has ancestors of all colors, we keep it.

Lemma 14. *The tree τ is a suffix tree iff there is a survived location of s.*

Proof. If there is such a location, then for any x in the suffix tour graph of the augmented tree we have $d(x) \geq 0$ and there is a cycle containing the root and all leaves. We are still to apply the local changes caused by implicit parents. Namely, for each node x with an implicit parent the edge from y to x is to be replaced by the path y, $par_{\$}(x)$, x (see Fig. 7b). The cycle can be re-routed to go via the new paths instead of the edges, and it will contain the root and the leaves of τ. Hence, the augmented tree is a \$-suffix tree and τ is a suffix tree.

If τ is a suffix tree, then it can be augmented to become a \$-suffix tree. The parent of s will survive the selection process. □

Suppose that there is such a location. Then we can find the parent of the twist node t if the twist node exists. It must have a child that satisfies $(\star)$, and we choose this child to be the twist node t. Let the first letter on the edge to the twist node be a. Then we put the first letter on all new leaf edges caused by the implicit nodes equal to a. The resulting graph will be the suffix tour graph of a \$-suffix tree. We can use the solution of I et. al [15] to reconstruct a string $S\$$ realizing this \$-suffix tree in linear time. The tree τ will be a suffix tree of the string S. This completes the proof of Theorem 2. □

5. Conclusions and Open Problems

We have proved several new properties of suffix trees, including an upper bound of $n-1$ on the length of a shortest string S realizing a suffix tree τ with n nodes. As noted this bound is tight in terms of n, since the number of leaves in τ, which can be $n-1$, provides a trivial lower bound on the length of S.

Using these properties, we have shown how to decide if a tree τ with n nodes is a suffix tree in $O(n)$ time, provided that the suffix links of internal nodes and

the first letter on each edge is specified. It remains an interesting open question whether the problem can be solved without first letters or, even, without suffix links (i.e., given only the tree structure).

Our results imply that the set of all \$-suffix trees is a proper subset of the set all of suffix trees (e.g., the suffix tree of a string *abaababaa* is not a \$-suffix tree (see Fig. 5 and Lemma 7), which in turn is a proper subset of the set of all trees (consider, e.g., Fig. 1c or simply a path of length 2).

References

[1] Alberto Apostolico, Maxime Crochemore, Martin Farach-Colton, Zvi Galil, and S. Muthukrishnan. Forty Years of Text Indexing. In *Proc. 24th CPM (LNCS 7922)*, pages 1–10. 2013.

[2] Hideo Bannai, Shunsuke Inenaga, Ayumi Shinohara, and Masayuki Takeda. Inferring strings from graphs and arrays. In *Proc. 28th MFCS (LNCS 2747)*, volume 2747, pages 208–217, 2003.

[3] Dany Breslauer and Ramesh Hariharan. Optimal parallel construction of minimal suffix and factor automata. *Parallel Process. Lett.*, 06(01):35–44, 1996.

[4] Dany Breslauer and Giuseppe F. Italiano. On suffix extensions in suffix trees. *Theor. Comp. Sci.*, 457:27–34, 2012.

[5] Julien Clément, Maxime Crochemore, and Giuseppina Rindone. Reverse engineering prefix tables. In *Proc. 26th STACS*, pages 289–300, 2009.

[6] Maxime Crochemore, Costas S. Iliopoulos, Solon P. Pissis, and German Tischler. Cover array string reconstruction. In *Proc. 21st CPM (LNCS 6129)*, pages 251–259, 2010.

[7] Jean-Pierre Duval, Thierry Lecroq, and Arnaud Lefebvre. Border array on bounded alphabet. *J. Autom. Lang. Comb.*, 10(1):51–60, 2005.

[8] Jean-Pierre Duval, Thierry Lecroq, and Arnaud Lefebvre. Efficient validation and construction of border arrays and validation of string matching automata. *RAIRO Theor. Inform. Appl.*, 43:281–297, 2009.

[9] Jean-Pierre Duval and Arnaud Lefebvre. Words over an ordered alphabet and suffix permutations. *RAIRO Theor. Inform. Appl.*, 36(3):249–259, 2002.

[10] Paweł Gawrychowski, Artur Jeż, and Łukasz Jeż. Validating the Knuth-Morris-Pratt failure function, fast and online. *Theory Comput. Syst.*, 54(2):337–372, 2014.

[11] D. Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology.* Cambridge University Press, 1997.

[12] Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Counting parameterized border arrays for a binary alphabet. In *Proc. LATA (LNCS 5457)*, pages 422–433, 2009.

[13] Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Verifying a Parameterized Border Array in $O(n^{1.5})$ Time. In *Proc. 21st CPM (LNCS 6129)*, pages 238–250, 2010.

[14] Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Verifying and enumerating parameterized border arrays. *Theor. Comput. Sci.*, 412(50):6959–6981, 2011.

[15] Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Inferring strings from suffix trees and links on a binary alphabet. *Discrete Appl. Math.*, 163:316–325, 2014.

[16] Gregory Kucherov, Lilla Tóthmérész, and Stéphane Vialette. On the combinatorics of suffix arrays. *Inf. Process. Lett.*, 113(22-24):915–920, 2013.

[17] Weilin Lu, P. J. Ryan, W. F. Smyth, Yu Sun, and Lu Yang. Verifying a border array in linear time. *J. Comb. Math. Comb. Comput.*, 42:223–236, 2002.

[18] Esko Ukkonen. On-line construction of suffix trees. *Algorithmica*, 14(3):249–260, 1995.

[19] Peter Weiner. Linear pattern matching algorithms. In *Proc. 14th FOCS (SWAT)*, pages 1–11, 1973.