

# Inferring Strings from Position Heaps in Linear Time

Koshiro Kumagai, Diptarama Hendrian, Ryo Yoshinaka, and Ayumi Shinohara

Tohoku University, Japan

## Abstract

Position heaps are index structures of text strings used for the string matching problem. They are rooted trees whose edges and nodes are labeled and numbered, respectively. This paper is concerned with variants of the inverse problem of position heap construction and gives linear-time algorithms for those problems. The basic problem is to restore a text string from a rooted tree with labeled edges and numbered nodes. In the variant problems, the input trees may miss edge labels or node numbers which we must restore as well.

## 1 Introduction

The string matching problem searches for occurrences of a pattern  $P$  in a text  $T$ . It has been widely studied for many years and many efficient algorithms have been proposed. Those techniques can be classified into mainly two approaches. The first one is to construct data structures from  $P$  by preprocessing  $P$ . For example, the Knuth-Morris-Pratt algorithm [18] constructs border arrays, the Boyer-Moore method [4] constructs suffix tables, and the Z-algorithm [15] constructs prefix tables which is the dual notion of suffix tables. The other approach is preprocessing  $T$  to create indexing structures, such as suffix trees [25], suffix arrays [20], LCP arrays [20], suffix graphs [2], compact suffix graphs [3], and position heaps [11]. Indexing structures are advantageous when searching for many different patterns in a text.

The reverse engineering of those data structures has also been widely studied. Studying reverse engineering deepens our insight into those data structures. For example, it may enable us to design an algorithm generating indexing structures with specific structural characteristics, which should be useful for verifying other software processing them. The early studies targeted border arrays [8, 9, 13]. Later, Clément et al. [7] proposed a linear time algorithm for inferring strings from prefix tables. Those data structures are produced by preprocessing patterns. The reverse engineering for indexing structures has been studied for suffix arrays [1, 10], LCP arrays [17], suffix graphs [1], and suffix trees [5, 16, 23]. The techniques used in [16] and [23] involve finding an Eulerian cycles on a graph modifying an input tree.

In this paper, we discuss the reverse engineering of another type of indexing structures, called *position heaps* [11, 19]. The position heap of a string  $T$  is a rooted tree with labeled edges and numbered nodes. Actually, Ehrenfeucht et al. [11] and Kucherov [19] gave different definitions of position heaps. By either definition, position heaps can be constructed in linear time online assuming the alphabet size to be constant. In addition, we can find all occurrence positions of a pattern  $P$  in  $O(|P|^2 + k)$  time, where  $k$  is the output size. Moreover, by augmenting position heaps with additional data structures, we can improve the searching time to  $O(|P| + k)$ .

We consider the following four types of reverse engineering of Kucherov's position heaps [19]. The first problem is to restore a source text  $T$  from an input edge-labeled and node-numbered rooted tree so that the input should be the position heap  $\text{PH}(T)$  of  $T$ . While this problem allows at most one solution, the other problems may have many possible solutions. In the second problem, input trees miss edge labels. In the third problem, input trees miss node numberings. Instance trees of the fourth problem miss both edge labels and node numberings but have potential *suffix links* among nodes,

which play an important role in the construction of position heaps. We show that all the problems above can be solved in linear time in the input size. Among those, we devote the most pages to the third problem. We reduce the problem to finding a special type of Eulerian cycle over the input tree augmented with suffix links. By showing the problem of finding an Eulerian cycle of this special type is linear-time solvable, we conclude that restoring a text from a position heap without node numbers is linear-time solvable. This can be seen analogous to the techniques used in [16] and [23] for the suffix tree reverse engineering. In addition, we present formulas for counting the number of possible text strings, which can be computed in polynomial time. Moreover, we show efficient algorithms for enumerating all possible text strings in output linear time.

## 2 Preliminaries

Let  $\Sigma$  be a finite alphabet and let the size of  $\Sigma$  be constant. For a string  $w$  over  $\Sigma$ , the length of  $w$  is denoted by  $|w|$ . The *empty string*  $\varepsilon$  is the string of length 0. Throughout this paper, strings are 1-indexed. For  $1 \leq i \leq j \leq |w|$ , we let  $w[i]$  be the  $i$ -th letter of  $w$ , and  $w[i : j]$  be the substring of  $w$  which starts at position  $i$  and ends at position  $j$ . In particular, we denote  $w[i : |w|]$  by  $w[i : ]$  and  $w[1 : j]$  by  $w[: j]$ . The concatenation of two strings  $s$  and  $t$  is denoted by  $st$ .

Let  $\mathbb{N}_0$  and  $\mathbb{N}_1$  be the set of natural numbers including and excluding 0, respectively. We denote the cardinality of a set  $X$  by  $|X|$ .

### 2.1 Graphs

A *directed multigraph*  $G$  is a tuple  $(V, E, \Gamma)$  where  $V$  is the node set,  $E \subseteq V \times V$  is the edge set, and  $\Gamma : E \rightarrow \mathbb{N}_1$  gives each edge its multiplicity. The *head* and the *tail* of an edge  $(u, v) \in E$  are  $v$  and  $u$ , respectively. This paper disallows self-loops:  $(v, v) \notin E$  for any  $v \in V$ . When  $\Gamma(e) = 1$  for all  $e \in E$ ,  $G$  is called a *directed graph* and is simply denoted by  $(V, E)$ . An *edge-labeled multigraph* is a tuple  $(V, E, \Gamma, \Psi)$  where  $\Psi : E \rightarrow \Sigma$  for an alphabet  $\Sigma$ . A sequence  $p = \langle e_1, \dots, e_\ell \rangle$  of edges is called a  $v_0$ - $v_\ell$  *path* if there are  $v_0, \dots, v_\ell \in V$  such that  $e_i = (v_{i-1}, v_i)$  for all  $i \in \{1, \dots, \ell\}$ . Note that, the same node may occur more than once in a path in this paper. We call  $p$  a  $v_0$ -*cycle* when  $v_0 = v_\ell$ . For a  $t$ - $u$  path  $p_1$  and a  $u$ - $v$  path  $p_2$ , we denote by  $p_1 \cdot p_2$  the concatenation of  $p_1$  and  $p_2$ , which will be a  $t$ - $v$  path. By extending the domain of  $\Psi$  to sequences of edges, we define the *path label*  $\Psi(p)$  of  $p$  to be the string  $\Psi(e_1) \cdots \Psi(e_\ell)$ . When there exists just one  $v_0$ - $v_\ell$  path, we call its label the  $v_0$ - $v_\ell$  path label and denote it by  $\Psi((v_0, v_\ell)) \in \Sigma^*$ .

A directed graph  $G$  is a  $t$ -*rooted tree* ( $t \in V$ ) if there exists exactly one  $t$ - $v$  path for all  $v \in V$ . We call  $t$  the *root* of  $G$ . Similarly,  $G$  is a  $t$ -*oriented tree* if there exists exactly one  $v$ - $t$  path for all  $v \in V$ . We call  $t$  the *sink* of  $G$ . For a  $t$ -rooted tree  $G = (V, E)$ , if  $(u, v) \in E$ , then  $u$  is the *parent* of  $v$  and  $v$  is a *child* of  $u$ . For two nodes  $u, v \in V$  such that a  $u$ - $v$  path exists,  $v$  is a *descendant* of  $u$ , and  $u$  is an *ancestor* of  $v$ . The *depth* of  $v$  is the length of the unique path from the root to  $v$ . We denote the set of all descendants of  $v$  as  $\mathcal{D}_G(v)$ .

Two directed multigraphs  $G = (V, E, \Gamma)$  and  $G' = (V', E', \Gamma')$  are *isomorphic*, denoted by  $G \equiv G'$ , if there is a bijection  $\phi$  over  $V$  such that  $V' = \phi(V)$ ,  $E' = \{(\phi(u), \phi(v)) \mid (u, v) \in E\}$ , and  $\Gamma'((\phi(u), \phi(v))) = \Gamma((u, v))$ . The definition of isomorphism is naturally extended and applied for edge-labeled directed multigraphs. When  $G$  is a rooted tree, we can verify  $G \equiv G'$  in linear time. If  $V' = V$  and  $G'$  is a  $t$ -oriented tree, then  $G'$  is a  $t$ -*oriented spanning tree* of  $G$ .

Let  $G = (V, E, \Gamma)$  be a directed multigraph. For a node  $v \in V$ ,  $\delta_G^-(v)$  and  $\delta_G^+(v)$  are the sets of edges whose heads and tails are  $v$ , respectively. We denote the sum of the multiplicities of edges contained in  $\delta_G^-(v)$  and  $\delta_G^+(v)$  by  $\Delta_G^-(v) = \sum_{e \in \delta_G^-(v)} \Gamma(e)$  and  $\Delta_G^+(v) = \sum_{e \in \delta_G^+(v)} \Gamma(e)$ , respectively. A cycle  $p$  is *Eulerian* when  $p$  contains  $e$  just  $\Gamma(e)$  times for all  $e \in E$ . We also call a directed multigraph *Eulerian* if it has an Eulerian cycle. It is well-known that  $G$  is Eulerian if and only if  $G$  is connected and  $\Delta_G^-(v) = \Delta_G^+(v)$  for all  $v \in V$  [12]. Therefore, we can check whether  $G$  is Eulerian in  $O(|V| + |E|)$  time. We often drop the subscript  $G$  from  $\mathcal{D}_G$ ,  $\delta_G^+$ ,  $\Delta_G^-$  etc. when  $G$  is clear from the context.

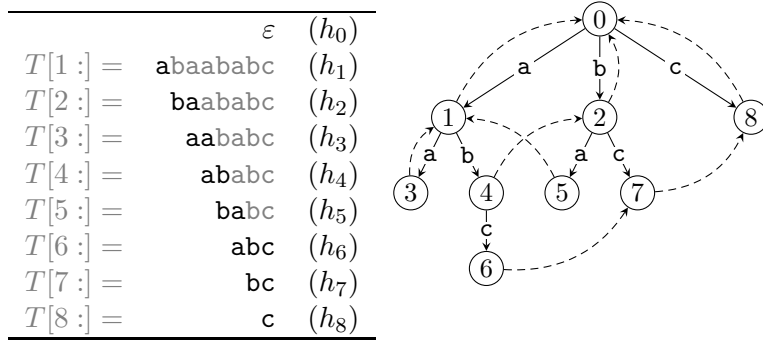


Figure 1:  $\text{PHS}(\text{abaababc})$  (dashed arrows are suffix links)

## 2.2 Position heaps

A position heap is an index structure with which one can efficiently solve the pattern matching problem. In this paper, we follow Kucherov’s definition [19]. Let  $T$  be a string of length  $n$  ending with a unique letter, i.e.,  $T[i] \neq T[n]$  for all  $i \in \{1, \dots, n-1\}$ . The position heap  $\text{PH}(T)$  of  $T$  is an edge-labeled rooted tree  $(V, E, \Psi)$  defined as follows. Let  $h_0$  be  $\varepsilon$ , and  $h_i$  be the shortest prefix of  $T[i:]$  not contained in  $\{h_0, \dots, h_{i-1}\}$  for all  $i \in \{1, \dots, n\}$ . Since  $T$  ends with a unique letter,  $T[i:] \neq h_j$  for any  $j < i$ , and thus  $h_i$  is always defined. Then, define  $V = \{0, \dots, n\}$ ,  $E = \{(i, j) \mid h_i c = h_j \text{ for some } c \in \Sigma\}$ , and  $\Psi((i, j)) = c$  if  $h_i c = h_j$ . Clearly, a position heap is 0-rooted and  $h_i$  is the 0- $i$  path label for all  $i \in \{0, \dots, n\}$ . Moreover, we have  $i \leq j$  if node  $i$  is an ancestor of node  $j$ . We call  $T$  the *source text* of  $\text{PH}(T)$ . Kucherov showed that one can determine whether a pattern  $P$  occurs in  $T$  in  $O(|P|^2)$  time using  $\text{PH}(T)$ . Moreover, we can determine it in  $O(|P|)$  time with auxiliary data structures.

In Kucherov’s algorithm for constructing position heaps, the mapping  $\mathcal{S}: V \setminus \{0\} \rightarrow V$  called *suffix links* plays an important role. It is defined by  $\mathcal{S}(i) = j$  such that  $h_i = ch_j$  for some  $c \in \Sigma$  for  $i > 0$ . The suffix links are well-defined. It is clear that the depth of node  $i$  is the depth of node  $\mathcal{S}(i)$  plus 1. We often treat  $\mathcal{S}$  as a subset of  $V \times V$ . We denote the position heap augmented with its suffix links by  $\text{PHS}(T) = (V, E, \Psi, \mathcal{S})$ . Figure 1 shows  $\text{PHS}(T)$  for  $T = \text{abaababc}$ .

## 2.3 Problem definitions

In this paper, we consider the following inverse problems of position heap construction. The first problem is inferring the source text  $T$  from a position heap.

**Problem 1** (Inferring source texts from node-numbered edge-labeled trees).

**Input:** An edge-labeled rooted tree  $(V, E, \Psi)$  with  $V = \{0, \dots, |V| - 1\}$ .

**Output:** A string  $T$  such that  $\text{PH}(T) = (V, E, \Psi)$  if such  $T$  exists. Otherwise, “invalid”.

We will also consider the problem where edge labels are missing.

**Problem 2** (Inferring source texts from node-numbered trees).

**Input:** A rooted tree  $(V, E)$  with  $V = \{0, \dots, |V| - 1\}$ .

**Output:** A string  $T$  such that  $\text{PH}(T) = (V, E, \Psi)$  for some  $\Psi$  if such  $T$  exists. Otherwise, “invalid”.

The third problem is inferring source texts  $T$  from trees whose nodes are not numbered but edges are labeled.

**Problem 3** (Inferring source texts from edge-labeled trees).

**Input:** An edge-labeled rooted tree  $(V, E, \Psi)$ .

**Output:** A string  $T$  such that  $\text{PH}(T) \equiv (V, E, \Psi)$  if such  $T$  exists. Otherwise, “invalid”.

In the end, we will address the problem where the input trees miss both node numbers and edge labels but have potential suffix links.

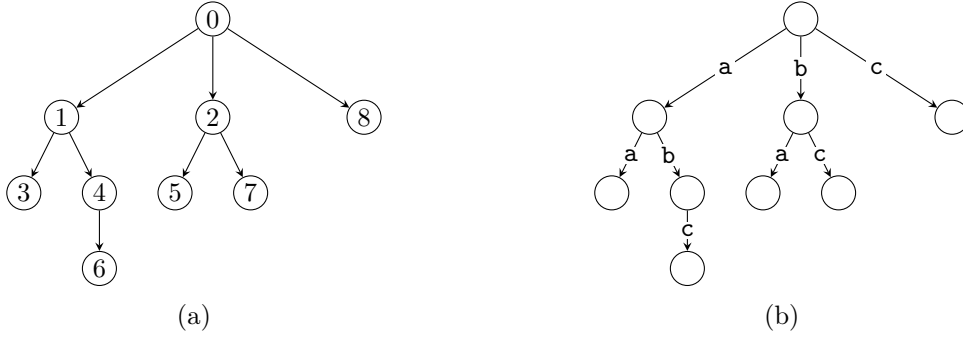


Figure 2: Examples of inputs to instances of (a) Problem 2 and (b) Problem 3.

**Problem 4** (Inferring source texts from trees with links).

**Input:** A pair  $(G, \mathcal{S})$  of a rooted tree  $G = (V, E)$  and a partial map  $\mathcal{S}: V \rightarrow V$ .

**Output:** A string  $T$  such that  $\text{PHS}(T) \equiv (V, E, \Psi, \mathcal{S})$  for some  $\Psi$  if such  $T$  exists. Otherwise, “invalid”.

Figure 2 shows examples of instances of Problem 2 and 3 and Figure 3 shows all possible answers for the instance of Figure 2(b).

### 3 Proposed algorithms

#### 3.1 Inferring source texts from node-numbered edge-labeled trees

Solving Problem 1 is easy. Given an edge-labeled tree  $(V, E, \Psi)$  where  $V = \{0, \dots, n\}$ , let  $h_i$  be the  $0$ - $i$  path label on  $G$  for every  $i \in V$ . If the input is the position heap of some string  $T$ , it must hold  $T[i] = h_i[1]$ . Therefore, by DFS on  $G$  remembering the initial letter of each path label, we can construct the candidate string  $T$  in linear time. Then, we can verify whether  $\text{PH}(T) = (V, E, \Psi)$  in linear time, since the position heap of  $T$  can be constructed in linear time [19].

**Theorem 1.** *Problem 1 is solvable in linear time.*

#### 3.2 Inferring source texts from node-numbered trees

Figure 2(a) shows an input to an instance of Problem 2. The following procedure solves Problem 2. We label the outgoing edges of the root with arbitrary but distinct letters of  $\Sigma$ . Then, we construct an output candidate  $T$  following the method for Problem 1 in the previous subsection.

**Theorem 2.** *Problem 2 is solvable in linear time.*

There can be many correct outputs for input unless it is invalid. The number of possible source texts to output equals the number of how to attach the labels to edges from the root  $r$ . Since the number of letters that appear in  $T$  equals  $\Delta^+(r)$ , the number of possible texts is  $|\Sigma|! / (|\Sigma| - \Delta^+(r))!$ . One can enumerate such  $T$  in output linear time because one can enumerate all  $\Delta^+(r)$ -permutations of  $\Sigma$  in output linear time [22].

#### 3.3 Inferring source texts from edge-labeled trees

Compared to the previous two problems, solving Problem 3 in linear time requires more elaborate arguments. In this subsection, we assume that two distinct outgoing edges of a node have different labels, since otherwise obviously the input cannot be extended to a position heap. We will investigate the structural properties of position heaps augmented with the suffix links, and see that the text  $T$  will appear as the label of a path with a specific property over  $\text{PHS}(T)$ .

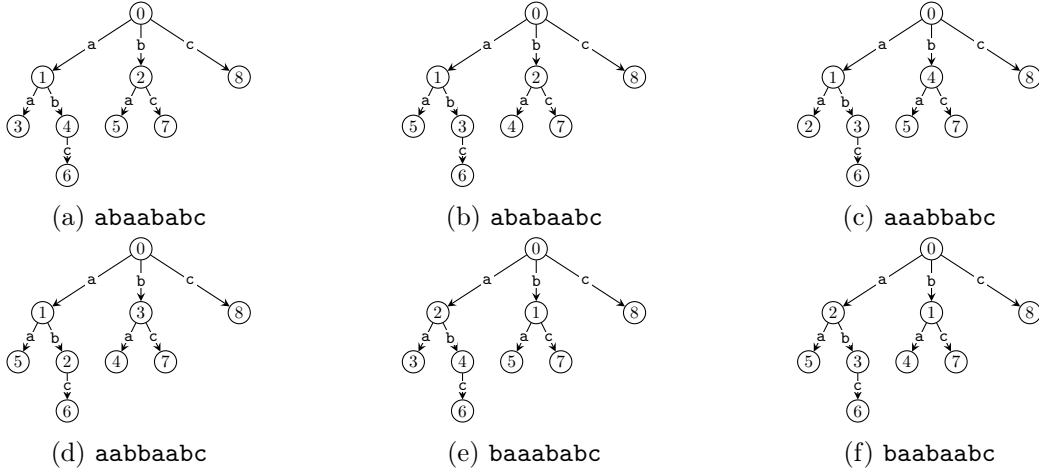


Figure 3: All possible answers to Problem 3 when the graph in Figure 2(b) is given.

**Lemma 1.** Let  $\text{PHS}(T) = (V, E, \Psi, \mathcal{S})$  with  $V = \{0, \dots, n\}$ . We have  $\mathcal{S}(n) = 0$  and  $i + 1 \in \mathcal{D}(\mathcal{S}(v))$  for all  $i \in V \setminus \{0, n\}$ .

*Proof.* We show the lemma by induction on the depth of node  $i$ . When the depth of node  $i$  is 1,  $\mathcal{S}(i)$  is the root 0, which is an ancestor of every node including  $i + 1$ . Note that the depth of node  $n$  is 1 since  $T$  ends with a unique letter. Suppose the depth of node  $i < n$  is two or more. In this case, let the  $0-i$  path label  $h_i$  be  $awb$  for some  $a, b \in \Sigma$  and  $w \in \Sigma^*$ . Let  $j$  be the parent of  $i$ , for which  $h_j = aw$ . Let  $i_s = \mathcal{S}(i)$  and  $j_s = \mathcal{S}(j)$ , i.e.,  $h_{i_s} = wb$  and  $h_{j_s} = w$ . By the induction hypothesis, we have  $j + 1 \in \mathcal{D}(j_s)$ , i.e.,  $h_{j+1} = h_{j_s}w'$  for some  $w' \in \Sigma^*$ , which implies that  $j + 1 \geq j_s$ . Together with the fact that  $i > j$ , we have  $i + 1 > j_s$ . Since  $h_i$  and  $h_{i+1}$  are prefixes of  $T[i : ]$  and  $T[i + 1 : ]$ , respectively, either  $h_i[2 : ] = wb$  is a prefix of  $h_{i+1}$  or the other way around. The fact  $h_{j_s} = w$  and  $i + 1 > j_s$  implies that  $wb$  is a prefix of  $h_{i+1}$ . That is,  $h_{i_s}$  is a prefix of  $h_{i+1}$ , which means  $i + 1 \in \mathcal{D}(i_s)$ .  $\square$

Hereafter, by a *path/cycle* of  $\text{PHS}(T) = (V, E, \Psi, \mathcal{S})$ , we mean a path/cycle of  $(V, E \cup \mathcal{S})$ . We call elements of  $E \cup \mathcal{S}$  *arcs* while reserving the term *edges* for elements of  $E$ . From Lemma 1, for all  $i \in \{1, \dots, n - 1\}$ ,  $\text{PHS}(T)$  has a special  $i-(i + 1)$  path which starts with the suffix link followed by zero or some number of edges. We define a cycle by concatenating all special  $i-(i + 1)$  paths.

**Definition 1.** For  $\text{PHS}(T) = (V, E, \Psi, \mathcal{S})$ , let  $f_i = (i, \mathcal{S}(i))$ ,  $p_0$  the path from 0 to 1, and  $p_i$  the path from  $\mathcal{S}(i)$  to  $i + 1$  for  $i > 0$ . The  $T$ -trace cycle of  $\text{PHS}(T)$  is the sequence  $p_0 \cdot f_1 \cdot p_1 \cdots f_{n-1} \cdot p_{n-1} \cdot f_n$ .

Figure 4 shows the  $T$ -trace cycle of  $\text{PHS}(T)$  for  $T = \text{abaababc}$ . Note that the  $T$ -trace cycle is a cycle in the graph  $(V, E \cup \mathcal{S})$ , where each element of  $\mathcal{S}$  appears exactly once. Since following an edge from  $E$  and a suffix link from  $\mathcal{S}$  increases and decreases the depth by one, respectively, the total numbers of occurrences of edges and suffix links in the  $T$ -trace cycle should be balanced. That is, the  $T$ -trace cycle contains exactly  $n$  occurrences of edges from  $E$ . The following lemma explains why we call the cycle  $T$ -trace cycle.

**Lemma 2.** Let  $e \in E$  be the  $i$ -th occurrence of an edge in the  $T$ -trace cycle of  $\text{PHS}(T)$ . Then  $\Psi(e) = T[i]$ .

*Proof.* Suppose the  $i$ -th edge  $e = (u, v)$  in the  $T$ -trace cycle  $p = p_0 \cdot f_1 \cdots p_{n-1} \cdot f_n$  occurs in the  $p_j$  segment. In other words,  $p$  can be written as  $p' \cdot (u, v) \cdot p''$ , where  $p'$  contains  $j$  suffix links and  $i - 1$  edges. Then, the depth of  $v$  is  $i - j$ . Moreover, the edge  $e$  is on the path from the root to the node  $j + 1$ , whose label is a prefix of  $T[j + 1 : ]$ . That is,  $\Psi(e)$  is the  $(i - j)$ -th letter of  $T[j + 1 : ]$ . Hence,  $\Psi(e) = T[(j + 1) + (i - j) - 1] = T[i]$ .  $\square$

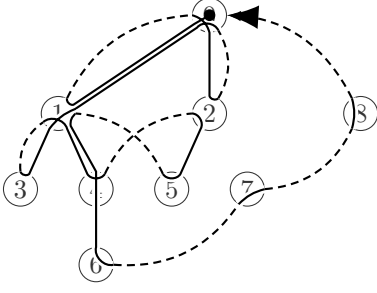


Figure 4: The  $T$ -trace cycle of  $\text{PHS}(T)$  with  $T = \text{abaababc}$ , which is an answer to the input graph in Figure 2(b). Dashed lines represent suffix links.

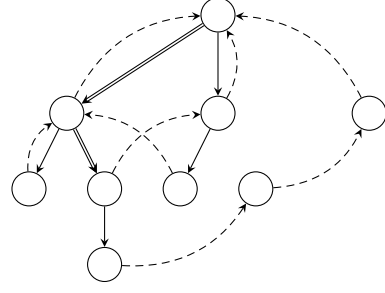


Figure 5: The trace graph of the input graph in Figure 2(b). The multiplicities of doubled edges are 2 and the others are 1. Dashed arrows show suffix links.

Lemma 2 allows us to spell  $T$  by following the  $T$ -trace cycle without referring to node numbers. To solve Problem 3, we will construct the  $T$ -trace cycle of  $\text{PHS}(T) \equiv (V, E, \Psi, \mathcal{S}_G)$  for some  $T$  from the input graph  $G = (V, E, \Psi)$ . For this end, we first reconstruct the suffix links  $\mathcal{S}$ .

**Lemma 3.** *From an edge-labeled rooted tree  $G = (V, E, \Psi)$ , one can uniquely construct  $\mathcal{S}$  in linear time such that  $\text{PHS}(T) \equiv (V, E, \Psi, \mathcal{S})$  for some  $T$  if any exist.*

*Proof.* We recover the suffix links of nodes from shallower to deeper. Let  $r$  be the root of  $G$ . From the definition of suffix links, we have  $\mathcal{S}(v) = r$  for every node of depth 1. For  $e = (u, v) \in E$  with  $\Psi(e) = c$ , we assume  $\mathcal{S}(u)$  has already been determined. Let  $aw$  be the  $r$ - $u$  path label where  $a \in \Sigma$  and  $w \in \Sigma^*$ . The  $r$ - $\mathcal{S}(u)$  path label is  $w$  and the  $r$ - $v$  path label is  $awc$ . Therefore, the  $r$ - $\mathcal{S}(v)$  path label is  $wc$ . Hence, an edge  $(\mathcal{S}(u), \mathcal{S}(v))$  labeled  $c$  exists. So, for the node  $t \in V$  such that  $(\mathcal{S}(u), t) \in E$  and  $\Psi((\mathcal{S}(u), t)) = c$ , we determine  $\mathcal{S}(v) = t$ .  $\square$

If we fail to give a suffix link to any of the nodes by the procedure described in the proof of Lemma 3, the answer to Problem 3 is “invalid”.

While the  $T$ -trace cycle contains just one occurrence of each suffix link, the numbers of occurrences of respective edges vary. Actually, one can uniquely determine the multiplicity of each edge in the  $T$ -trace cycle from  $G$ .

**Lemma 4.** *Let  $\sigma(e)$  be the number of occurrences of  $e$  in the  $T$ -trace cycle for all  $e \in E$ . Then, it holds that*

$$\sigma(e) = 1 - |\{u \in V \mid \mathcal{S}_G(u) = v\}| + \sum_{e' \in \delta_G^+(v)} \sigma(e') \quad (1)$$

where  $v$  is the head of  $e$ .

Note that  $\delta_G^+(v)$  contains no suffix links of  $\text{PHS}(T)$ .

*Proof.* The  $T$ -trace cycle must include the same number of occurrences of arcs coming into and going out from node  $v$ . Since each suffix link occurs just once in the  $T$ -trace cycle, we obtain the lemma.  $\square$

**Lemma 5.** *The system of equations (1) in  $\sigma$  has a unique solution. Moreover, it can be computed in linear time.*

*Proof.* One can uniquely determine the value of  $\sigma(e)$  inductively on the height of  $e \in E$ . Then, the linear-time computation is obvious.  $\square$

Let us call a cycle  $p$  of  $(V, E, \Psi, \mathcal{S})$  a *legitimate cycle* if it is the  $T$ -trace cycle for some  $T$ . Based on Lemmas 4 and 5, we define the directed multigraph for which every legitimate cycle is Eulerian.

**Definition 2** (Trace graph). The trace graph  $\mathcal{G}(G)$  of an edge-labeled tree  $G = (V, E, \Psi)$  is a tuple  $(V, E', \mathcal{S}, \Gamma)$  where  $E' = \{e \in E \mid \sigma(e) > 0\}$  and  $\Gamma: E' \cup \mathcal{S} \rightarrow \mathbb{N}_1$  is defined by

$$\Gamma(e) = \begin{cases} 1 & \text{if } e \in \mathcal{S}, \\ \sigma(e) & \text{if } e \in E', \end{cases}$$

where  $\mathcal{S}$  and  $\sigma$  are given in Lemmas 3 and 5, respectively.

Figure 5 shows the trace graph of Figure 2(b). The doubled arrows have multiplicity 2 and the others have 1. The dashed arrows are suffix links.

From the definition, it is obvious that the  $T$ -trace cycle is an  $r$ -Eulerian cycle of  $\mathcal{G}(G)$  where  $r$  is the root of  $G$ . However, not every Eulerian cycle of  $\mathcal{G}(G)$  can be a legitimate cycle. Recall that in the definition of the  $T$ -trace cycle, the suffix link of every node  $u$  proceeds all outgoing edges of  $u$ . We say that an Eulerian cycle  $p$  of  $\mathcal{G}(G)$  *respects*  $\mathcal{S}$  if no edges of  $\delta_G^+(u)$  occur before  $(u, \mathcal{S}(u))$  in  $p$ .

**Lemma 6.** A cycle  $p$  is an  $r$ -Eulerian cycle respecting  $\mathcal{S}$  if and only if  $p$  is the  $T$ -trace cycle of some  $T$ .

*Proof.* ( $\Leftarrow$ ) By definition.

( $\Rightarrow$ ) Let  $n = |V|$  and  $r$  be the root of  $G$ . Let  $p_i$  and  $f_{i+1}$  be the sequences of edges and the suffix links for  $i = 0, \dots, n-1$  so that  $p = p_0 \cdot f_1 \cdot p_1 \dots p_{n-1} \cdot f_n$ . Since  $p$  ends at  $r$  and only suffix links point to  $r$ ,  $p$  always ends with a suffix link. We define the bijection  $\Lambda: V \rightarrow \{0, \dots, n\}$  such that  $\Lambda(r) = 0$  and  $\Lambda(s) = i$  if  $f_i = (s, \mathcal{S}(s))$  for all  $s \in V \setminus \{r\}$ . Let  $s_i$  be the node such that  $\Lambda(s_i) = i$ .

We first show  $\Lambda(u) < \Lambda(v)$  for all  $(u, v) \in E$  by induction on  $\Lambda(v)$ . Suppose the claim holds true for all  $v$  such that  $\Lambda(v) < i$ . Then, we will show the claim holds for the edge whose head is  $s_i$ . If  $|p_i| \geq 1$ , the edge  $(s_k, s_i)$  occurs just before  $f_i = (s_i, \mathcal{S}(s_i))$  in  $p$ . Since  $p$  respects  $\mathcal{S}$ ,  $f_k = (s_k, \mathcal{S}(s_k))$  occurs before  $(s_k, s_i)$ . Thus, we have  $k < i$ . If  $|p_i| = 0$ ,  $f_{i-1} = (s_{i-1}, s_i)$ . Let the parents of  $s_{i-1}$  and  $s_i$  be  $s_j$  and  $s_k$ , respectively. By the induction hypothesis,  $j < i-1$ . By the definition of  $\mathcal{S}$ ,  $f_j = (s_j, s_k) \in \mathcal{S}$ . Since  $p$  respects  $\mathcal{S}$ ,  $f_k = (s_k, \mathcal{S}(s_k))$  appears either before  $f_j$  or right after  $f_j$ . That is,  $k \leq j+1$  holds. Therefore,  $k < i$ .

Now, we define a string  $T$  by  $T[i] = \Psi(e_i)$  where  $e_i$  is the  $i$ -th edge in  $p$  for  $i = 1, \dots, n$ , and define  $h_i$  inductively to be the shortest prefix of  $T[i:]$  which is not in  $\{h_0, \dots, h_{i-1}\}$  where  $h_0 = \varepsilon$ . We will show by induction on  $i$  that for all  $j \leq i$ , the  $s_0$ - $s_j$  path label  $\Psi((s_0, s_j))$  is  $h_j = T[j : x_j]$  where  $x_j = |p_0 \dots p_{j-1}|$ . This implies  $(V, E, \Psi) \equiv \text{PH}(T)$  when  $i = n$ . Then the constructed  $\mathcal{S}$  is the correct suffix links of  $\text{PH}(T)$  by Lemma 3 and thus  $p$  is the  $T$ -trace cycle.

Let  $g_i = \Psi((s_0, s_i))$ . The claim clearly holds for  $i = 0$  by  $g_0 = h_0 = \varepsilon$ . Suppose the claim holds true for  $i$ . That is,  $g_i = h_i = T[i : x_i]$  where  $x_i = |p_0 \dots p_{i-1}|$ . Let  $u = \mathcal{S}(s_i)$ . By the definition of  $\mathcal{S}$ , we have  $\Psi((s_0, u)) = g_i[2:] = T[i+1 : x_i]$ . By the definition of  $T$ ,  $\Psi((u, s_{i+1})) = p_i = T[x_i+1 : x_i+|p_i|] = T[x_i+1 : x_{i+1}]$ , where  $x_{i+1} = |p_0 \dots p_i|$ . By concatenating these two paths, we obtain  $g_{i+1} = \Psi((s_0, s_{i+1})) = T[i+1 : x_{i+1}]$ . Since the labels of all proper ancestors of  $s_{i+1}$  are at most  $i$ , all prefixes of  $g_{i+1}$  appears in  $\{g_0, \dots, g_i\} = \{h_0, \dots, h_i\}$ . That is,  $g_{i+1}$  is the least prefix of  $T[i+1:]$  not in  $\{h_0, \dots, h_i\}$ , i.e.,  $g_{i+1} = h_{i+1}$ .  $\square$

Therefore, to find a source text  $T$ , it is enough to find an  $r$ -Eulerian cycle over  $(V, E, \mathcal{S}, \Gamma)$  that respects  $\mathcal{S}$  where  $r$  is the root. We show that this problem can be solved in linear time on general graphs.

**Problem 5** (The ECP (Eulerian cycle with priority edges) problem).

**Input:** A tuple  $(G, F, r)$  of a directed multigraph  $G = (V, E, \Gamma)$ , an edge subset  $F \subseteq E$ , and a start node  $r \in V$  such that  $|F \cap \delta^+(v)| \leq 1$  for all  $v \in V$  and  $\Gamma(e) = 1$  for all  $e \in F$ .

**Output:** An  $r$ -Eulerian cycle that respects  $F$  if any. Otherwise, “invalid”.

We call edges of  $F$  *priority edges*. Without loss of generality, we may assume a node has a priority outgoing edge only if it has another outgoing edge. If a node has only one outgoing edge and it has

priority, then one can remove it from  $F$  and make it a non-priority edge. This does not affect possible solutions. In what follows, we show how to solve the ECP problem in linear time.

First, let us review a linear-time algorithm for constructing an  $r$ -Eulerian cycle. The following procedure gives a justification for the so-called BEST theorem [24, 6], which counts the number of Eulerian cycles in a directed multigraph.

1. Construct an arbitrary  $r$ -oriented spanning tree  $H$  of  $G$ ,
2. Starting from  $r$ , choose an arbitrary unused edge to follow next, except that an edge in  $H$  can be chosen only when it is the only remaining choice, until we follow all the edges of  $G$ .

This process guarantees to find an Eulerian cycle without getting stuck. We modify this procedure so that the output shall respect  $F$ .

1. Construct an arbitrary  $t$ -oriented spanning tree  $H$  of  $(V, E \setminus F)$ ,
2. Starting from  $r$ , choose an arbitrary unused edge to follow next, except that
  - choose an unused priority edge if the current node has any,
  - an edge in  $H$  can be chosen only when it is the only remaining choice,
until we follow all the edges of  $G$ .

**Theorem 3.** *We can compute an answer to the ECP problem in linear time.*

One can count the number of  $r$ -ECPs by modifying the BEST theorem formula. Letting  $G' = (V, E \setminus F, \Gamma')$  with the restriction  $\Gamma'$  of  $\Gamma$  to  $E \setminus F$ , the number of  $r$ -ECPs is given as

$$\Delta_{G'}^+(r) \cdot \prod_{v \in V} \frac{(\Delta_{G'}^+(v) - 1)!}{\prod_{e \in \delta_{G'}^+(v)} \Gamma'(e)!} \cdot \sum_{(V, E') \in \mathcal{T}_{G'}(r)} \prod_{e \in E'} \Gamma'(e) \quad (2)$$

where  $\mathcal{T}_{G'}(r)$  is the set of  $r$ -oriented spanning trees of  $G'$ . One can compute (2) in polynomial time by the matrix-tree theorem [21].

**Theorem 4.** *We can calculate the number of  $r$ -ECPs in polynomial time.*

One can also enumerate  $r$ -ECPs. We have already described a linear-time nondeterministic algorithm to find an  $r$ -ECP. Gabow and Myers proposed an algorithm [14] to enumerate spanning trees in output linear time. By searching all the possible choices of the procedure, we enumerate all the  $r$ -ECPs.

**Theorem 5.** *We can enumerate  $r$ -ECPs in linear time per solution.*

**Corollary 1.** *Problem 3 is solvable in linear time. Moreover, one can count and enumerate all possible answers in polynomial time and output linear time, respectively.*

*Proof.* The first claim follows from Theorem 3. By Theorems 4 and 5, it suffices to show that two distinct legitimate cycles  $p$  and  $p'$  over a trace graph give different source texts. Suppose  $e$  and  $e'$  are the first mismatch of  $p$  and  $p'$ . Since choosing a suffix link is obligatory,  $e \neq e'$  implies  $e, e' \in E$ . Since distinct edges with the same tail have distinct labels,  $\Psi(e) \neq \Psi(e')$ , and thus those two cycles spell different source texts.  $\square$

### 3.4 Inferring source texts from trees with links

Instance trees of Problem 4 miss both node numbers and edge labels but have possible suffix links. This problem can be solved by combining ideas for solving Problems 2 and 3. We first label the outgoing edges of the root node with arbitrary distinct letters. Then, the other edge labels are uniquely determined by the definition of suffix links, as long as the input is valid. Now, the algorithm for Problem 3 can be applied. Similarly one can solve the counting and enumerating variants of Problem 4.

**Theorem 6.** *We can solve Problem 4 in linear time. Moreover, one can count the number of output strings in polynomial time, and enumerate all output strings in linear time per each.*



## 4 Conclusion

We studied four types of reverse engineering problems on Kucherov’s position heaps [19] and showed that all problems can be solved in linear time. One can think of an even more restrictive variant, where the input tree has no edge labels, no node numbers, and no suffix links. In this setting, we need to find “valid” suffix links, which seems a challenging task.

One can also study the reverse engineering problems of position heaps based on the definition by Ehrenfeucht et al. [11]. We conjecture that those problems can be solved by quite similar techniques presented in this paper.

Another interesting direction of future work is to study the reverse engineering of augmented position heaps [11].

## References

- [1] Hideo Bannai, Shunsuke Inenaga, Ayumi Shinohara, and Masayuki Takeda. Inferring strings from graphs and arrays. In *Proc. MFCS 2003*, pages 208–217, 2003.
- [2] Anselm Blumer, Janet Blumer, David Haussler, Andrzej Ehrenfeucht, Mu-Tian Chen, and Joel Seiferas. The smallest automaton recognizing the subwords of a text. *Theoretical Computer Science*, 40:31–55, 1985.
- [3] Anselm Blumer, Janet Blumer, David Haussler, Ross McConnell, and Andrzej Ehrenfeucht. Complete inverted files for efficient text retrieval and analysis. *Journal of the ACM*, 34(3):578–595, 1987.
- [4] Robert S. Boyer and J. Strother Moore. A fast string searching algorithm. *Communications of the ACM*, 20:762–772, 1977.
- [5] Bastien Cazaux and Eric Rivals. Reverse engineering of compact suffix trees and links: A novel algorithm. *Journal of Discrete Algorithms*, 28:9–22, 2014.
- [6] Charalambos A. Charalambides. *Enumerative combinatorics*, volume 2. Chapman and Hall/CRC, 2018.
- [7] Julien Clément, Maxime Crochemore, and Giuseppina Rindone. Reverse engineering prefix tables. In *Proc. STACS 2009*, pages 289–300, 2009.
- [8] Jean-Pierre Duval, Thierry Lecroq, and Arnaud Lefebvre. Border array on bounded alphabet. *Journal of Automata, Languages and Combinatorics*, 10(1):51–60, 2005.
- [9] Jean-Pierre Duval, Thierry Lecroq, and Arnaud Lefebvre. Efficient validation and construction of border arrays and validation of string matching automata. *RAIRO-Theoretical Informatics and Applications*, 43(2):281–297, 2009.
- [10] Jean-Pierre Duval and Arnaud Lefebvre. Words over an ordered alphabet and suffix permutations. *RAIRO-Theoretical Informatics and Applications*, 36(3):249–259, 2002.
- [11] Andrzej Ehrenfeucht, Ross M. McConnell, Nissa Osheim, and Sung-Whan Woo. Position heaps: A simple and dynamic text indexing data structure. *Journal of Discrete Algorithms*, 9(1):100–121, 2011.
- [12] Herbert Fleischner. *Eulerian graphs and related topics*, volume 1. Elsevier, 1990.
- [13] Frantisek Franek, Weilin Lu, P J Ryan, William F Smyth, Yu Sun, and Lu Yang. Verifying a border array in linear time. *Journal on Combinatorial Mathematics and Combinatorial Computing*, 42:223–236, 2002.

- [14] Harold N. Gabow and Eugene W. Myers. Finding all spanning trees of directed and undirected graphs. *SIAM Journal on Computing*, 7(3):280–287, 1978.
- [15] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.
- [16] Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Inferring strings from suffix trees and links on a binary alphabet. *Discrete Applied Mathematics*, 163:316–325, 2014.
- [17] Juha Kärkkäinen, Marcin Piatkowski, and Simon J. Puglisi. String inference from longest-common-prefix array. In *Proc. ICALP 2017*, pages 62:1–62:14, 2017.
- [18] Donald E. Knuth, Jr. James H. Morris, and Vaughan R. Pratt. Fast string searching in strings. *SIAM Journal on Computing*, 6(2):323–350, 1977.
- [19] Gregory Kucherov. On-line construction of position heaps. *Journal of Discrete Algorithms*, 20:3 – 11, 2013.
- [20] Udi Manber and Gene Myers. Suffix arrays: A new method for on-line string searches. *SIAM Journal on Computing*, 22(5):935–948, 1993.
- [21] Cristopher Moore and Stephan Mertens. *The nature of computation*. OUP Oxford, 2011.
- [22] Robert Sedgewick. Permutation generation methods. *ACM Computing Surveys (CSUR)*, 9(2):137–164, 1977.
- [23] Tatiana Starikovskaya and Hjalte Wedel Vildhøj. A suffix tree or not a suffix tree? *Journal of Discrete Algorithms*, 32:14–23, 2015.
- [24] T. van Aardenne-Ehrenfest and N. G. de Bruijn. Circuits and trees in oriented linear graphs. *Simon Stevin : Wis- en Natuurkundig Tijdschrift*, 28:203–217, 1951.
- [25] P. Weiner. Linear pattern matching algorithm. In *Proc. 14th IEEE Symposium on Switching and Automata Theory*, pages 1–11, 1973.