


DAWGs for Parameterized Matching: Online Construction and Related Indexing Structures

Katsuhito Nakashima

Graduate School of Information Sciences,
Tohoku University, Sendai, Japan
katsuhito_nakashima@shino.ecei.tohoku.ac.jp

Diptarama Hendrian 


Graduate School of Information Sciences,
Tohoku University, Sendai, Japan
diptarama@tohoku.ac.jp

Ryo Yoshinaka 

Graduate School of Information Sciences,
Tohoku University, Sendai, Japan
ryoshinaka@tohoku.ac.jp

Hideo Bannai 

M&D Data Science Center,
Tokyo Medical and Dental University,
Tokyo, Japan
hdbn.dsc@tmd.ac.jp

Masayuki Takeda 

Department of Informatics,
Kyushu University, Fukuoka, Japan
takeda@inf.kyushu-u.ac.jp

Noriki Fujisato

Department of Informatics,
Kyushu University, Fukuoka, Japan
noriki.fujisato@inf.kyushu-u.ac.jp

Yuto Nakashima 

Department of Informatics,
Kyushu University, Fukuoka, Japan
yuto.nakashima@inf.kyushu-u.ac.jp

Shunsuke Inenaga 

Department of Informatics,
Kyushu University, Fukuoka, Japan
PRESTO, Japan Science and Technology Agency,
Kawaguchi, Japan
inenaga@inf.kyushu-u.ac.jp

Ayumi Shinohara 

Graduate School of Information Sciences,
Tohoku University, Sendai, Japan
ayumis@tohoku.ac.jp

Abstract

Two strings x and y over $\Sigma \cup \Pi$ of equal length are said to *parameterized match* (p -match) if there is a renaming bijection $f : \Sigma \cup \Pi \rightarrow \Sigma \cup \Pi$ that is identity on Σ and transforms x to y (or vice versa). The p -matching problem is to look for substrings in a text that p -match a given pattern. In this paper, we propose *parameterized suffix automata* (p -suffix automata) and *parameterized directed acyclic word graphs* (PDAWGs) which are the p -matching versions of suffix automata and DAWGs. While suffix automata and DAWGs are equivalent for standard strings, we show that p -suffix automata can have $\Theta(n^2)$ nodes and edges but PDAWGs have only $O(n)$ nodes and edges, where n is the length of an input string. We also give $O(n|\Pi| \log(|\Pi| + |\Sigma|))$ -time $O(n)$ -space algorithm that builds the PDAWG in a left-to-right online manner. As a byproduct, it is shown that the *parameterized suffix tree* for the reversed string can also be built in the same time and space, in a right-to-left online manner.

2012 ACM Subject Classification Theory of computation \rightarrow Pattern matching

Keywords and phrases parameterized matching, suffix trees, DAWGs, suffix automata

Digital Object Identifier 10.4230/LIPIcs.CPM.2020.26

Related Version A full version of the paper is available at [17], <https://arxiv.org/abs/2002.06786>.

Funding *Diptarama Hendrian*: JSPS KAKENHI Grant Number JP19K20208.

Yuto Nakashima: JSPS KAKENHI Grant Number JP18K18002.

Ryo Yoshinaka: JSPS KAKENHI Grant Number JP18H04091.



© Katsuhito Nakashima, Noriki Fujisato, Diptarama Hendrian, Yuto Nakashima, Ryo Yoshinaka, Shunsuke Inenaga, Hideo Bannai, Ayumi Shinohara, and Masayuki Takeda; licensed under Creative Commons License CC-BY

31st Annual Symposium on Combinatorial Pattern Matching (CPM 2020).

Editors: Inge Li Gørtz and Oren Weimann; Article No. 26; pp. 26:1–26:14



Leibniz International Proceedings in Informatics

LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Shunsuke Inenaga: JSPS KAKENHI Grant Number JP17H01697, JST PRESTO Grant Number JPMJPR1922.

Hideo Bannai: JSPS KAKENHI Grant Numbers JP16H02783, JP20H04141.

Ayumi Shinohara: JSPS KAKENHI Grant Number JP15H05706.

Masayuki Takeda: JSPS KAKENHI Grant Number JP18H04098.

1 Introduction

The *parameterized matching problem* (*p-matching problem*) [2] is a class of pattern matching where the task is to locate substrings of a text that have “the same structure” as a given pattern. More formally, we consider a parameterized string (p-string) over a union of two disjoint alphabets Σ and Π for static characters and for parameter characters, respectively. Two equal length p-strings x and y are said to *parameterized match* (*p-match*) if x can be transformed to y (and vice versa) by a bijection which renames the parameter characters. The *p-matching problem* is, given a text p-string T and pattern p-string P , to report the occurrences of substrings of T that p-match P . P-matching is well-motivated by plagiarism detection, software maintenance, RNA structural pattern matching, and so on [2, 18, 15, 16].

The *parameterized suffix tree* (*p-suffix tree*) [1] is the fundamental indexing structure for p-matching, which supports p-matching queries in $O(m \log(|\Pi| + |\Sigma|) + pocc)$ time, where m is the length of pattern P , and $pocc$ is the number of occurrences to report. It is known that the p-suffix tree of a text w of length n can be built in $O(n \log(|\Pi| + |\Sigma|))$ time with $O(n)$ space in an offline manner [13] and in a *left-to-right online* manner [18]. A *randomized* $O(n)$ -time left-to-right online construction algorithm for p-suffix trees is also known [14]. Indexing p-strings has recently attracted much attention, and the p-matching versions of other indexing structures, such as *parameterized suffix arrays* [6, 12, 3, 9], *parameterized BWTs* [11], and *parameterized position heaps* [7, 8, 10], have also been proposed.

This paper fills in the missing pieces of indexing structures for p-matching, by proposing the parameterized version of the *directed acyclic word graphs* (DAWGs) [4, 5], which we call the *parameterized directed acyclic word graphs* (PDAWGs).

For any standard string T , the three following data structures are known to be equivalent:

- (1) The *suffix automaton* of T , which is the minimum DFA that is obtained by merging isomorphic subtrees of the suffix trie of T .
- (2) The DAWG, which is the edge-labeled DAG of which each node corresponds to a equivalence class of substrings of T defined by the set of ending positions in T .
- (3) The *Weiner-link graph*, which is the DAG consisting of the nodes of the suffix tree of the reversal \bar{T} of T and the reversed suffix links (a.k.a. soft and hard Weiner links).

The equality of (2) and (3) in turn implies symmetry of suffix trees and DAWGs, namely:

- (a) The suffix links of the DAWG for T form the suffix tree for \bar{T} .
- (b) Left-to-right online construction of the DAWG for T is equivalent to right-to-left online construction of the suffix tree for \bar{T} .

Firstly, we present (somewhat surprising) combinatorial results on the p-matching versions of data structures (1) and (2). We show that the *parameterized suffix automaton* (*p-suffix automaton*), which is obtained by merging isomorphic subtrees of the *parameterized suffix trie* of a p-string T of length n , can have $\Theta(n^2)$ nodes and edges in the worst case, while the PDAWG for any p-string has $O(n)$ nodes and edges. On the other hand, the p-matching versions of data structures (2) and (3) are equivalent: The *parameterized Weiner-link graph* of the p-suffix tree for \bar{T} is equivalent to the PDAWG for T . As a corollary to this, symmetry (a) also holds: The suffix links of the PDAWG for T form the p-suffix tree for \bar{T} .

Secondly, we present algorithmic results on PDAWG construction. We first propose left-to-right online construction of PDAWGs that works in $O(n|\Pi| \log(|\Pi| + |\Sigma|))$ time with $O(n)$ space. In addition, as a byproduct of this algorithm, we obtain a right-to-left online construction of the p-suffix tree in $O(n|\Pi| \log(|\Pi| + |\Sigma|))$ time with $O(n)$ space. This can be seen as the p-matching version of symmetry (b). We suspect that it is difficult to shave the $n|\Pi|$ term in the left-to-right online construction of PDAWGs, as well as in the right-to-left construction of p-suffix trees.

A full version of this work can be found in [17].

2 Preliminaries

We denote the set of all non-negative integers by \mathcal{N} . A linear order \prec over \mathcal{N} is identical to the ordinary linear order $<$ on integers except that 0 is always bigger than any other positive integers: $a \prec b$ if and only if $0 < a < b$ or $a \neq b = 0$. For a nonempty finite subset S of \mathcal{N} , $\max_{\prec} S$ and $\min_{\prec} S$ denote the maximum and minimum elements of S with respect to the order \prec , respectively.

We denote the set of strings over an alphabet A by A^* . For a string $w = xyz \in A^*$, x , y , and z are called *prefix*, *factor*, and *suffix* of w , respectively. The sets of the prefixes, factors, and suffixes of a string w are denoted by $\text{Prefix}(w)$, $\text{Factor}(w)$, and $\text{Suffix}(w)$, respectively. The length of w is denoted by $|w|$ and the i -th symbol of w is denoted by $w[i]$ for $1 \leq i \leq |w|$. The factor of w that begins at position i and ends at position j is $w[i : j]$ for $1 \leq i \leq j \leq |w|$. For convenience, we abbreviate $w[1 : i]$ to $w[: i]$ and $w[i : |w|]$ to $w[i :]$ for $1 \leq i \leq |w|$. The empty string is denoted by ε , that is $|\varepsilon| = 0$. Moreover, let $w[i : j] = \varepsilon$ if $i > j$. The *reverse* \bar{w} of $w \in A^*$ is inductively defined by $\bar{\varepsilon} = \varepsilon$ and $\bar{a\bar{x}} = a\bar{x}$ for $a \in A$ and $x \in A^*$.

Throughout this paper, we fix two alphabets Σ and Π . We call elements of Σ *static* symbols and those of Π *parameter* symbols. Elements of Σ^* and $(\Sigma \cup \Pi)^*$ are called *static strings* and *parameterized strings* (or *p-strings* for short), respectively.

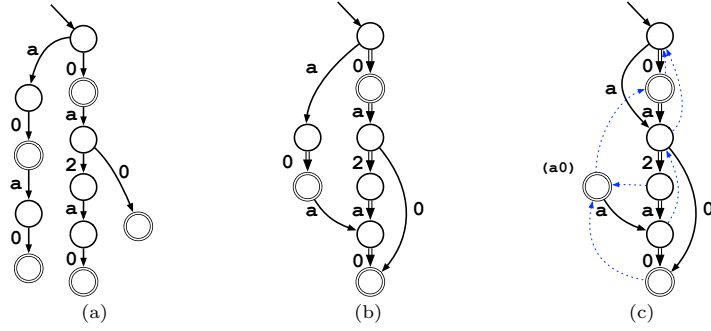
Given two p-strings S_1 and S_2 of length n , S_1 and S_2 are a *parameterized match* (*p-match*), denoted by $S_1 \approx S_2$, if there is a bijection f on $\Sigma \cup \Pi$ such that $f(a) = a$ for any $a \in \Sigma$ and $f(S_1[i]) = S_2[i]$ for all $1 \leq i \leq n$ [2]. The *prev-encoding* $\text{prev}(S)$ of a p-string S is the string over $\Sigma \cup \mathcal{N}$ of length $|S|$ defined by

$$\text{prev}(S)[i] = \begin{cases} S[i] & \text{if } S[i] \in \Sigma, \\ 0 & \text{if } S[i] \in \Pi \text{ and } S[i] \neq S[j] \text{ for } 1 \leq j < i, \\ i - j & \text{if } S[i] = S[j] \in \Pi, j < i \text{ and } S[i] \neq S[k] \text{ for any } j < k < i \end{cases}$$

for $i \in \{1, \dots, |S|\}$. We call a string $x \in (\Sigma \cup \mathcal{N})^*$ a *pv-string* if $x = \text{prev}(S)$ for some p-string S . For any p-strings S_1 and S_2 , $S_1 \approx S_2$ if and only if $\text{prev}(S_1) = \text{prev}(S_2)$ [2]. For example, given $\Sigma = \{\mathbf{a}, \mathbf{b}\}$ and $\Pi = \{\mathbf{u}, \mathbf{v}, \mathbf{x}, \mathbf{y}\}$, $S_1 = \mathbf{uvvauvb}$ and $S_2 = \mathbf{xyyaxyb}$ are a p-match by f such that $f(\mathbf{u}) = \mathbf{x}$ and $f(\mathbf{v}) = \mathbf{y}$, where $\text{prev}(S_1) = \text{prev}(S_2) = 001\mathbf{a}43\mathbf{b}$. For a p-string T , let $\text{PFactor}(T) = \{\text{prev}(S) \mid S \in \text{Factor}(T)\}$ and $\text{PSuffix}(T) = \{\text{prev}(S) \mid S \in \text{Suffix}(T)\}$ be the sets of prev-encoded factors and suffixes of T , respectively. For a factor $x \in (\Sigma \cup \mathcal{N})^*$ of a pv-string, the *re-encoding* $\langle x \rangle$ of x is the pv-string of length $|x|$ defined by $\langle x \rangle[i] = Z(x[i], i - 1)$ for $i \in \{1, \dots, |x|\}$ where

$$Z(a, j) = \begin{cases} 0 & \text{if } a \in \mathcal{N} \text{ and } a > j, \\ a & \text{otherwise.} \end{cases}$$

We then have $\langle \text{prev}(T)[i : j] \rangle = \text{prev}(T[i : j])$ for any i, j . We apply PFactor etc. to pv-strings w so that $\text{PFactor}(w) = \{\langle x \rangle \mid x \in \text{Factor}(w)\}$.



■ **Figure 1** (a) The parameterized suffix trie $\text{PSTrie}(T)$, (b) the parameterized suffix automaton $\text{PSAuto}(T)$ and (c) the PDAWG $\text{PDAWG}(T)$ for $T = \text{xaxay}$ over $\Sigma = \{\mathbf{a}\}$ and $\Pi = \{\mathbf{x}, \mathbf{y}\}$, for which $\text{prev}(T) = 0\mathbf{a}2\mathbf{a}0$. Solid and broken arrows represent the edges and suffix links, respectively. Some nodes of $\text{PDAWG}(T)$ cannot be reached by following the edges from the source node.

Let $w, x, y \in (\Sigma \cup \mathcal{N})^*$. A symbol $a \in \Sigma \cup \mathcal{N}$ is said to be a *right extension* of x w.r.t. w if $xa \in \text{PFactor}(w)$. The set of the right extensions of x is denoted by $\text{REx}_w(x)$. The set of the *end positions* of x in a pv-string w is defined by $\text{RPos}_w(x) = \{i \in \{0, \dots, |w|\} \mid x = \langle w[i - |x| + 1 : i] \rangle\}$. Note that $0 \in \text{RPos}_w(x)$ iff $x = \varepsilon$. It is easy to see that $\text{RPos}_w(x) \subseteq \text{RPos}_w(y)$ if and only if $y \in \text{PSuffix}(x)$ or $\text{RPos}_w(x) = \text{RPos}_w(y)$. We write $x \equiv_w^R y$ iff $\text{RPos}_w(x) = \text{RPos}_w(y)$, and the equivalence class of pv-strings w.r.t. \equiv_w^R as $[x]_w^R$. Note that for any $x \notin \text{PFactor}(w)$, $[x]_w^R$ is the infinite set of all the pv-strings outside $\text{PFactor}(w)$. For a finite nonempty set X of strings which has no distinct elements of equal length, the shortest and longest elements of X are denoted by $\min X$ and $\max X$, respectively.

A basic indexing structure of a p-strings is a *parameterized suffix trie*. The parameterized suffix trie $\text{PSTrie}(T)$ is the trie for $\text{PSuffix}(T)$. That is, $\text{PSTrie}(T)$ is a tree (V, E) whose node set is $V = \text{PSuffix}(T)$ and edge set is $E = \{(x, a, xa) \in V \times (\Sigma \cup \mathcal{N}) \times V\}$. An example can be found in Figure 1 (a). Like the standard suffix tries for static strings, the size of $\text{PSTrie}(T)$ can be $\Theta(|T|^2)$. Obviously we can check whether T has a substring that p-matches P of length m in $O(m \log(|\Sigma| + |\Pi|))$ time using $\text{PSTrie}(T)$, assuming that finding the edge to traverse for a given character takes $O(\log(|\Sigma| + |\Pi|))$ time by, e.g., using balanced trees. We use the same assumption on other indexing structures considered in this paper.

3 Parameterized DAWG

3.1 Parameterized suffix automata

One natural idea to define the parameterized counterpart of DAWGs for p-strings, which we actually do not take, is to merge isomorphic subtrees of parameterized suffix tries. In other words, the parameterized suffix automaton of T , denoted by $\text{PSAuto}(T)$, is the minimal deterministic finite automaton that accepts $\text{PSuffix}(T)$. Figure 1 (b) shows an example of a parameterized suffix automaton. However, the size of $\text{PSAuto}(T)$ can be $\Theta(|T|^2)$, as witnessed by a p-string $T_k = \mathbf{x}_1 \mathbf{a}_1 \dots \mathbf{x}_k \mathbf{a}_k \mathbf{x}_1 \mathbf{a}_1 \dots \mathbf{x}_k \mathbf{a}_k$ over $\Sigma = \{\mathbf{a}_1, \dots, \mathbf{a}_k\}$ and $\Pi = \{\mathbf{x}_1, \dots, \mathbf{x}_k\}$.

3.2 Parameterized directed acyclic word graphs

In this section, we present a new indexing structure for parameterized strings, which we call *parameterized directed acyclic word graphs (PDAWGs)*. A PDAWG can be obtained from a parameterized suffix trie by merging nodes whose ending positions are the same.

In the example of Figure 1 (a), the subtrees rooted at \mathbf{a} and $0\mathbf{a}$ have different shapes but $\text{RPos}_w(\mathbf{a}) = \text{RPos}_w(0\mathbf{a}) = \{2, 4\}$. Particularly, the 0-edges of those two nodes point at nodes $\mathbf{a}0$ and $0\mathbf{a}0$ with different ending position sets, which shall not be merged. Our solution to this obvious conflict is to use only the edges of the “representative” node among ones with the same ending position sets. In the example, we take out-going edges of $0\mathbf{a}$ and do not care those of \mathbf{a} . The resultant PDAWG by our solution is shown in Figure 1 (c). This might first appear nonsense: by reading $\mathbf{a}0$, whose ending positions are 3 and 5, one will reach to the sink node, whose ending position is 5, and consequently one will reach no node by reading $\mathbf{a}0\mathbf{a} \in \text{PFactor}(T)$. We will argue in the next subsection that still we can correctly perform parameterized matching using our PDAWG by presenting a p-matching algorithm.

► **Definition 1** (Parameterized directed acyclic word graphs). *Let $w = \text{prev}(T)$ for a parameterized text $T \in (\Sigma \cup \Pi)^*$. The parameterized directed acyclic word graph (PDAWG) $\text{PDAWG}(T) = \text{PDAWG}(w)$ of T is the directed acyclic graph (V_w, E_w) where*

$$V_w = \{ [x]_w^R \mid x \in \text{PFactor}(w) \},$$

$$E_w = \{ ([x]_w^R, c, [y]_w^R) \in V \times (\Sigma \cup \mathcal{N}) \times V \mid y = \max[x]_w^R \cdot c \}$$

together with suffix links

$$\text{SL}_w([x]_w^R) = [\langle y[2 : |y|] \rangle_w^R \text{ where } y = \min[x]_w^R.$$

The nodes $[\varepsilon]_w^R$ and $[w]_w^R$ are called the source and the sink, respectively. Suffix links are defined on non-source nodes.

PDAWGs have the same size bound as DAWGs, shown by Blumer et al. [4].

► **Theorem 2.** *PDAWG(T) has at most $2n - 1$ nodes and $3n - 4$ edges when $n = |T| \geq 3$. Those bounds are tight.*

By definition, a node u has an out-going edge labeled with a if and only if $a \in \text{REx}_w(\max u)$. For $a \in \text{REx}_w(\max u)$, by $\text{child}_w(u, a)$ we denote the unique element v such that $(u, a, v) \in E_w$. For $a \notin \text{REx}_w(\max u)$, we define $\text{child}_w(u, a) = \text{Null}$. For any $u \in V_w$, $\text{RPos}_w(\text{SL}_w(u))$ is the least proper superset of $\text{RPos}_w(u)$. The reversed suffix links form a tree with root $[\varepsilon]_w^R$. Actually, the tree is isomorphic to the parameterized suffix tree [2] for \bar{T} . We discuss the duality between PDAWGs and parameterized suffix tree in more detail in Subsection 3.5.

3.3 Parameterized pattern matching with PDAWGs

This subsection discusses how we can perform p-matching using our PDAWGs: We must reach a node $[x]_w^R \in V_w$ by reading $x \in \text{PFactor}(w)$ and reach no node if $x \notin \text{PFactor}(w)$. In DAWGs for static strings, by following the a -edge of $[x]_w^R$, we will arrive in $[xa]_w^R$, which is guaranteed by the fact that $x \equiv_w^R y$ implies $xa \equiv_w^R ya$. However, this does not hold for pv-strings. For instance, for $w = \text{prev}(\mathbf{xaxay}) = 0\mathbf{a}2\mathbf{a}0$ ($\mathbf{a} \in \Sigma$ and $\mathbf{x}, \mathbf{y} \in \Pi$), we see $\text{RPos}_w(\mathbf{a}) = \text{RPos}_w(0\mathbf{a}) = \{2, 4\}$ but $3 \in \text{RPos}_w(\mathbf{a}0) \setminus \text{RPos}_w(0\mathbf{a}0)$. Consequently $\mathbf{a}0\mathbf{a} \in \text{PFactor}(w)$ but $0\mathbf{a}0\mathbf{a} \notin \text{PFactor}(w)$. By definition, if we reach a node u by reading $\max u$, we can simply follow the a -edge by reading a symbol a , similarly to matching using a DAWG. We may behave differently after we have reached u by reading some other string in u . The following lemma suggests how we can perform p-matching using $\text{PDAWG}(T)$.

■ **Algorithm 1** Parameterized pattern matching algorithm based on PDAWG(T).

```

1  $p \leftarrow \text{prev}(P)$ ;
2 Let  $u \leftarrow [\varepsilon]_w^R$ ;
3 for  $i = 1$  to  $|P|$  do
4   Let  $u \leftarrow \text{Trans}(u, i - 1, p[i])$ ;
5   if  $u = \text{Null}$  then return False;
6 return True;
```

■ **Algorithm 2** Function $\text{Trans}(u, i, a)$.

```

1 if  $a \neq 0$  then return  $\text{child}(u, a)$ ;
2 else if there is no  $b \in \text{rex}(u)$  such that  $b \succ i$  then return Null;
3 else if there is only one  $b \in \text{rex}(u)$  such that  $b \succ i$  then return  $\text{child}(u, b)$ ;
4 else return  $\text{SL}(\text{child}(u, b))$  for the smallest (w.r.t.  $\prec$ )  $b \in \text{rex}(u)$  such that  $b \succ i$ ;
```

► **Lemma 3.** Suppose $x \in \text{PFactor}(w)$ and $a \in \Sigma \cup \mathcal{N}$. Then, for $y = \max[x]_w^R$,

$$[xa]_w^R = \begin{cases} [ya]_w^R & \text{if } a \neq 0 \text{ or } W = \emptyset, \\ [yk]_w^R & \text{if } a = 0 \text{ and } |W| = 1, \\ \text{SL}_w([yk]_w^R) & \text{if } a = 0 \text{ and } |W| \geq 2, \end{cases}$$

where $W = \{j \in \mathcal{N} \mid yj \in \text{PFactor}(w) \text{ and } j \succ |x|\}$ and $k = \min_{\prec} W$.

Proof. We first show for $x \in \text{PFactor}(w)$, $a \in \Sigma \cup \mathcal{N}$ and $y = \max[x]_w^R$,

$$\text{RPos}_w(xa) = \begin{cases} \bigcup_{k \in W} \text{RPos}_w(yk) & \text{if } a = 0, \\ \text{RPos}_w(ya) & \text{otherwise.} \end{cases} \quad (1)$$

where $W = \{k \in \mathcal{N} \mid yk \in \text{PFactor}(w) \text{ and } k \succ |x|\}$.

For $a \in \Sigma$, $i \in \text{RPos}_w(xa)$ iff both $i - 1 \in \text{RPos}_w(x) = \text{RPos}_w(y)$ and $T[i] = a$ hold iff $i \in \text{RPos}_w(ya)$. For $a \in \mathcal{N} \setminus \{0\}$, noting that $0 < a \leq |x|$, we have $i \in \text{RPos}_w(xa)$ iff

$$i - 1 \in \text{RPos}_w(x) = \text{RPos}_w(y), T[i] = T[i - a] \in \Pi \text{ and } T[i - b] \neq T[i] \text{ for all } 0 < b < a$$

iff $i \in \text{RPos}_w(ya)$. For $a = 0$, $i \in \text{RPos}_w(xa)$ iff

$$i - 1 \in \text{RPos}_w(x) = \text{RPos}_w(y), T[i] \in \Pi, \text{ and } T[i] \neq T[j] \text{ for all } i - |x| < j < i$$

iff $i \in \text{RPos}_w(yk)$ for some $k \succ |x|$. This proves Eq. (1).

If $a \neq 0$ or $|W| \leq 1$, we obtain the lemma immediately from Eq. (1). Suppose $a = 0$ and $|W| \geq 2$ and let $k = \min_{\prec} W$. By Eq. (1), we see that $\text{RPos}_w(yk) \subsetneq \text{RPos}_w(x0)$, where $k \succ |x|$. It is enough to show that for any z , $\text{RPos}_w(yk) \subseteq \text{RPos}_w(z)$ implies either $\text{RPos}_w(yk) = \text{RPos}_w(z)$ or $\text{RPos}_w(x0) \subseteq \text{RPos}_w(z)$. Since $\text{RPos}_w(yk) \subseteq \text{RPos}_w(z)$, $z \in \text{PSuffix}(yk)$. If $z = z'k$, by $|x| < k < |z'| \leq |y|$, $z \in [yk]_w^R$. Then, Eq. (1) implies $\text{RPos}_w(z'k) = \text{RPos}_w(yk)$. Suppose $z = z'0$ for some z' . If $|z| \leq |x0|$, $\text{RPos}_w(x0) \subseteq \text{RPos}_w(z)$. Otherwise, $|x0| < |z| < |yk|$ implies $z' \in [y]_w^R$. By Eq. (1), $|z'| < k$ implies $\text{RPos}_w(x0) \subseteq \text{RPos}_w(z)$ by the choice of k . ◀

The function Trans of Algorithm 2 is a straightforward realization of Lemma 3. By $\text{rex}(u)$ we denote the set of labels of the out-going edges of u , i.e., $\text{rex}(u) = \text{REx}_w(\max u)$. It takes a node $u \in V$, a natural number $i \in \mathcal{N}$, and a symbol $a \in \Sigma \cup \mathcal{N}$, and returns the node

where we should go by reading a from u assuming that we have read i symbols so far. That is, $\text{Trans}([x]_w^R, |x|, a) = [xa]_w^R$ for every $xa \in \text{PFactor}(w)$. Using Trans , Algorithm 1 performs p-matching. We can locate the node v of the PDAWG in $O(m \log(|\Sigma| + |\Pi|))$ for a given pattern P of length m if it has a p-matching occurrence, or can determine that P does not have such an occurrence. In case P has a p-matching occurrence, we can actually report all of its occurrences by traversing the subtree of the (reversed) suffix links that is rooted at the node v , since the reversed suffix link tree of $\text{PDAWG}(T)$ forms the p-suffix tree of \bar{T} (see Subsection 3.5). Thus we obtain the following:

► **Theorem 4.** *Using $\text{PDAWG}(T)$ enhanced with the suffix links, we can find all substrings of T that p-match a given pattern P in $O(m \log(|\Sigma| + |\Pi|) + \text{pocc})$ time, where m is the length of pattern P and pocc is the number of occurrences to report.*

3.4 Online algorithm for constructing PDAWGs

This subsection proposes an algorithm constructing the PDAWG online. Our algorithm is based on the one by Blumer et al. [4] for constructing DAWGs of static strings. We consider updating $\text{PDAWG}(w)$ to $\text{PDAWG}(wa)$ for a pv-string wa where $a \in \Sigma \cup \mathcal{N}$.

We first observe properties similar to the DAWG construction.

► **Definition 5.** *The longest repeated suffix (LRS) of a nonempty pv-string $wa \in (\Sigma \cup \mathcal{N})^+$ is defined to be $\text{LRS}(wa) = \max(\text{PSuffix}(wa) \cap \text{PFactor}(w))$. If $\text{LRS}(wa) \neq \varepsilon$, the string obtained from $\text{LRS}(wa)$ by removing the last symbol is called the pre-LRS w.r.t. wa and denoted as $\text{preLRS}(wa) = \text{LRS}(wa)[: |\text{LRS}(wa)| - 1]$.*

Note that the pre-LRS w.r.t. wa is a suffix of w and is defined only when $\text{LRS}(wa) \neq \varepsilon$. We have $\text{LRS}(wa) = \varepsilon$ if and only if a is new in the sense that $wa \in \Sigma^* \{0\} \cup (\Sigma \cup \mathcal{N} \setminus \{a\})^* \Sigma$.

The following lemma for node splits on PDAWGs is an analogue to that for DAWGs.

► **Lemma 6 (Node update).** *For $x = \text{LRS}(wa)$ and $y = \max[x]_w^R$,*

$$V_{wa} = V_w \setminus \{[x]_w^R\} \cup \{[x]_{wa}^R, [y]_{wa}^R, [wa]_{wa}^R\}.$$

If $x = y$, then $[x]_w^R = [x]_{wa}^R = [y]_{wa}^R$, i.e., $V_{wa} = V_w \cup \{[wa]_{wa}^R\}$. Otherwise, $[x]_w^R = [x]_{wa}^R \cup [y]_{wa}^R$ and $[x]_{wa}^R \neq [y]_{wa}^R$.

Proof. First remark that $\text{RPos}_{wa}(z) = \text{RPos}_w(z) \cup \{|wa|\}$ for all $z \in \text{PSuffix}(wa)$ and $\text{RPos}_{wa}(z) = \text{RPos}_w(z)$ for all $z \notin \text{PSuffix}(wa)$. For those $z \in \text{PSuffix}(wa) \setminus \text{PFactor}(w)$, we have $\text{RPos}_{wa}(z) = \{|wa|\}$ and $[wa]_{wa}^R = \text{PSuffix}(wa) \setminus \text{PFactor}(w) \in V_{wa} \setminus V_w$. For $z \in \text{PFactor}(w)$, if $[z]_w^R \neq [z]_{wa}^R$, some elements of $[z]_w^R$ are in $\text{PSuffix}(wa)$ and some are not. That is, $[z]_w^R$ is partitioned into two non-empty equivalence classes $\{z' \in [z]_w^R \mid z' \in \text{PSuffix}(wa)\}$ and $\{z' \in [z]_w^R \mid z' \notin \text{PSuffix}(wa)\}$. By definition, the longest of the former is $x = \text{LRS}(wa)$ and the longest of the latter is $y = \max[x]_w^R$. Otherwise, $[z]_w^R = [z]_{wa}^R \in V_w \cap V_{wa}$. ◀

► **Example 7.** Let $w = 0a2a$ and $a = 0$. Then $\text{LRS}(wa) = \langle w[2 : 3] \rangle = \langle wa[4 : 5] \rangle = a0$. We have $\text{LRS}(wa) \neq \max[\text{LRS}(wa)]_w^R = 0a2$, where $\text{RPos}_w(a0) = \text{RPos}_w(0a2) = \{3\}$. On the other hand, $\text{RPos}_{wa}(a0) = \{3, 5\} \neq \text{RPos}_{wa}(0a2) = \{3\}$. Therefore, $\text{PDAWG}(wa)$ has two more nodes than $\text{PDAWG}(w)$.

When updating $\text{PDAWG}(w)$ to $\text{PDAWG}(wa)$, all edges that do not involve the node $[\text{LRS}(wa)]_w^R$ are kept by definition. What we have to do is to manipulate in-coming edges for the new sink node $[wa]_{wa}^R$, and, if necessary, to split the LRS node $[\text{LRS}(wa)]_w^R$ into two

and to manipulate in-coming and out-going edges of them. Therefore, it is very important to identify the LRS node $[\text{LRS}(wa)]_w^R$ and to decide whether $\text{LRS}(wa) = \max[\text{LRS}(wa)]_w^R$. The special case where $\text{LRS}(wa) = \varepsilon$ is easy to handle, since the LRS node will never be split by $[\varepsilon]_w^R = \{\varepsilon\}$. Hereafter we assume that $\text{LRS}(wa) \neq \varepsilon$ and $\text{preLRS}(wa)$ is defined. The LRS node can be reached from the pre-LRS node $[\text{preLRS}(wa)]_w^R$, which can be found by following suffix links from the sink node $[w]_w^R$ of $\text{PDAWG}(w)$. This appears quite similar to online construction of DAWGs for static strings, but there are nontrivial differences. Main differences from the DAWG construction are in the following points:

- Our PDAWG construction uses $\text{Trans}_w(u, i, Z(a, i))$ with an appropriate i , when the original DAWG construction refers to $\text{child}_w(u, a)$,
- While $\text{preLRS}(wa)$ is the longest of its equivalence class for static strings in $\text{DAWG}(w)$, it is not necessarily the case for p-strings (like the one in Figure 1), which affects the procedure to find the node of $\text{LRS}(wa)$,
- When a node of $\text{PDAWG}(w)$ is split into two in $\text{PDAWG}(wa)$, the out-going edges of the two nodes are identical in the DAWG construction, while it is not necessarily the case any more in our PDAWG construction. Moreover, we do not always have an edge from the node of $\text{preLRS}(wa)$ to that of $\text{LRS}(wa)$ in $\text{PDAWG}(wa)$.

In DAWGs, the pre-LRS node is the first node with an a -edge that can be found by recursively following the suffix links from the old sink $[w]_w^R$. However, it is not necessarily the case for PDAWGs. The following lemma suggests how to find $[\text{LRS}(wa)]_w^R$ and $|\text{LRS}(wa)|$ and how to decide whether the node shall be split.

► **Lemma 8.** *Let $x' = \text{preLRS}(wa)$, $a' = Z(a, |x'|)$, i.e., $x'a' = \text{LRS}(wa)$, and $u_i = \text{SL}_w^i(w)$ for $i \geq 0$.*

1. *We have $x' \in u_i$ for the least i such that $\text{Trans}_w(u_i, |\min u_i|, Z(a, |\min u_i|)) \neq \text{Null}$,*
2. $|x'a'| = \begin{cases} |\max[x']_w^R| + 1 & \text{if } a \in \text{REx}_w(\max[x']_w^R), \\ \min_{\prec}\{a, \max(\text{REx}(\max[x']_w^R) \cap \mathcal{N})\} & \text{otherwise,} \end{cases}$
3. $[x'a']_w^R = \text{Trans}_w(u_i, |x'|, a')$,
4. $[x'a']_w^R \neq [x'a']_{wa}^R$ if and only if $|x'a'| \neq |\max[x'a']_w^R|$.

Proof. Suppose $x' \in u_i$.

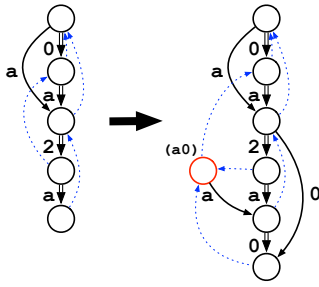
(1) Every string $z \in u_j$ with $j < i$ is properly longer than x' , so $z \cdot Z(a, |z|) \notin \text{PFactor}(w)$ by definition. On the other hand, for $z = \min u_i$, the fact $z \cdot Z(a, |z|) \in \text{PFactor}(x'a')$ implies $\text{Trans}_w(u_i, |z|, Z(a, |z|)) \neq \text{Null}$.

(2) If $a \in \text{REx}_w(y')$ for $y' = \max[x']_w^R$, we have $y'a \in \text{PFactor}(w)$ and thus $y'a = x'a'$. Suppose $a \notin \text{REx}_w(y')$. In this case, $[x']_w^R$ is not a singleton and thus not the source node, i.e., $|x'| \neq 0$. We have $\text{Trans}_w(u_i, |x'|, a') \neq \text{Trans}_w(u_i, |x'| + 1, Z(a, |x'| + 1)) = \text{Null}$ and thus $a \in \mathcal{N}$. Let $W = \text{REx}_w(\max u_i) \cap \mathcal{N}$ and $W_j = \{k \in W \mid k \succ j\}$. Lemma 3 implies that $a' = 0$ and $W_{|x'|} \neq \emptyset$ by $\text{Trans}_w(u_i, |x'|, a') \neq \text{Null}$. If $W_{|x'|+1} \neq \emptyset$, then $Z(a, |x'| + 1) = a \neq 0 = Z(a, |x'|)$, i.e., $|x'| = a - 1$. By $W_{|x'|} \neq \emptyset$, $|x'| = a - 1 \prec \max_{\prec} W_{|x'|}$. Therefore, $|x'| = \min_{\prec}\{a, \max_{\prec} W\} - 1$. If $W_{|x'|+1} = \emptyset \neq W_{|x'|}$, then $0 \notin W$ and $\max_{\prec} W = |x'| + 1$. By $Z(a, |x'|) = 0$, $|x'| \prec a$. Therefore, $|x'| = \min_{\prec}\{a, \max_{\prec} W\} - 1$.

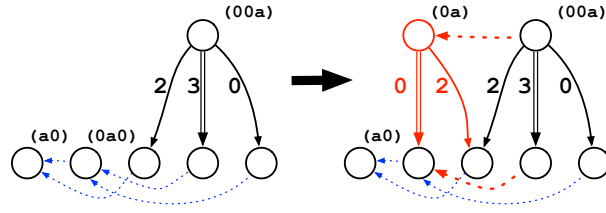
(3) By Lemma 3. (4) By Lemma 6. ◀

Edges are created or replaced in accordance with the definition of a PDAWG. The in-coming edges for the new sink node $[wa]_{wa}^R$ of $\text{PDAWG}(wa)$ are given as follows.

► **Lemma 9 (In-coming edges of the new sink).** *If $\text{LRS}(wa) \neq \varepsilon$, the in-coming edges for the new sink $[wa]_{wa}^R$ are exactly those $(u, Z(a, |\max u|), [wa]_{wa}^R)$ such that $u = \text{SL}_w^i([w]_w^R)$ for some $i \geq 0$ and $\text{child}(u, Z(a, |\max u|)) = \text{Null}$, i.e., $|\max u| > |\text{preLRS}(wa)|$. If $\text{LRS}(wa) = \varepsilon$, the in-coming edges for $[wa]_{wa}^R$ are exactly those $(\text{SL}_w^i([w]_w^R), a, [wa]_{wa}^R)$ for all $i \geq 0$.*



■ **Figure 2** PDAWG(w), PDAWG(wa) for $w = 0a2a$, $a = 0$. $LRS(wa) = x = a0$, $preLRS(wa) = x' = a$ and $y = \max[x]_w^R = 0a2$.



■ **Figure 3** Parts of PDAWG(w) and PDAWG(wa) for $w = 00a30a20a0$, $a = a$. $[0a]_{wa}^R$ does not inherit the out-going edges of $[0a]_w^R$ labeled with 3 and 0. Instead, the 3-edge and 0-edge are bundled into a single 0-edge which points at $Trans_w([0a]_w^R, 2, 0) = SL([00a3]_w^R) = [0a0]_w^R$.

This is not much different from DAWG update, except that the pre-LRS node has an edge towards the new sink when the pre-LRS is not the longest in the pre-LRS node (see Figure 2, where the pre-LRS node $[x']_{wa}^R$ has got a 0-edge towards the sink). If the LRS node $[LRS(wa)]_{wa}^R$ is not split, we have nothing more to do on edges.

Hereafter, we suppose that the LRS node must be split. That is, $x \neq y$ for $x = LRS(wa)$ and $y = \max[LRS(wa)]_w^R$. By definition, all edges of PDAWG(w) that do not involve the LRS node $[LRS(wa)]_w^R$ will be inherited to PDAWG(wa). The nodes $[x]_{wa}^R$ and $[y]_{wa}^R$ in PDAWG(wa) will have the following in-coming and out-going edges.

► **Lemma 10** (In-coming edges of the LRS node). *We have*

- $(u, b, [y]_{wa}^R) \in E_{wa}$ if and only if $(u, b, [y]_w^R) \in E_w$ and $|\max u| + 1 > |x|$,
- $(u, b, [x]_{wa}^R) \in E_{wa}$ if and only if $b = Z(a, |\max u|)$, $(u, b, [y]_w^R) \in E_w$ and $|\max u| + 1 \leq |x|$.

Lemma 10 is no more than a direct implication of the definition of edges of PDAWGs. An important fact is that $(u, b, [y]_w^R) \in E_w$ only if $u = SL_w^i([x']_w^R)$ with $x' = preLRS(wa)$ for some $i \geq 0$, which is essentially no difference from the DAWG case. Therefore, one can find all in-coming edges that may need to manipulate by following suffix links from the pre-LRS node. Note that in the on-line construction of a DAWG, the edge from the pre-LRS node $[x']_w^R$ to the LRS node $[y]_w^R$ in the old DAWG will be inherited to the new node $[x]_{wa}^R$ in the new DAWG. However, it is not necessarily the case in the PDAWG construction, as demonstrated in Figure 2, where the 2-edge from $[x']_w^R$ to $[y]_w^R$ in PDAWG(w) is kept as the 2-edge from $[x']_{wa}^R$ to $[y]_{wa}^R$ in PDAWG(wa) and, as a result, the new node $[x]_{wa}^R$ has no in-coming edges.

► **Lemma 11** (Out-going edges of the LRS node). *We have*

- $([y]_{wa}^R, b, u) \in E_{wa}$ if and only if $([y]_w^R, b, u) \in E_w$,
- $([x]_{wa}^R, b, u) \in E_{wa}$ if and only if $Trans([y]_w^R, |x|, b) = u$ if and only if either $([y]_w^R, b, u) \in E_w$ and $Z(b, |x|) \neq 0$ or $Trans([y]_w^R, |x|, 0) = u$ and $Z(b, |x|) = 0$.

Lemma 11 is also an immediate consequence of the definition of PDAWG edges. In the DAWG construction, those two nodes $[x]_{wa}^R$ and $[y]_{wa}^R$ simply inherit the out-going edges of the LRS node $[x]_w^R = [y]_w^R$. However, in the PDAWG construction, due to the prev-encoding rule on variable symbols, the node $[x]_{wa}^R$ will lose edges whose labels are integers greater than $|x|$, as demonstrated in Figure 3. Those edges are “bundled” into a single 0-edge which points at $Trans([y]_w^R, 0, |x|)$.

Updates of suffix links simply follow the definition.

■ **Algorithm 3** Constructing PDAWG(T).

```

1 Let  $V \leftarrow \{\top, \rho\}$ ,  $E \leftarrow \{(\top, a, \rho) \mid a \in \Sigma \cup \{0\}\}$ ,  $\text{SL}(\rho) = \top$ ,  $\text{len}(\top) = -1$ ,  $\text{len}(\rho) = 0$ ,
    $\text{sink} \leftarrow \rho$ , and  $t \leftarrow \text{prev}(T)$ ;
2 for  $i \leftarrow 1$  to  $|t|$  do
3   Let  $a \leftarrow t[i]$  and  $u \leftarrow \text{sink}$ ;
4   Create a new node and let  $\text{sink}$  be that node with  $\text{len}(\text{sink}) = i$ ;
5   while  $\text{Trans}(u, \text{len}(\text{SL}(u)) + 1, Z(a, \text{len}(\text{SL}(u)) + 1)) = \text{Null}$  do
6     Let  $\text{child}(u, Z(a, \text{len}(u))) \leftarrow \text{sink}$  and  $u \leftarrow \text{SL}(u)$ ;
   //  $u$  corresponds to  $[\text{preLRS}(t[:i])]_{t[:i-1]}^R$ 
7   if  $Z(a, \text{len}(u)) \in \text{rex}(u)$  then //  $\text{preLRS}(t[:i]) = \max[\text{preLRS}(t[:i])]_{t[:i-1]}^R$ 
8     Let  $k \leftarrow \text{len}(u) + 1$  and  $v \leftarrow \text{child}(u, Z(a, \text{len}(u)))$ 
9   else //  $\text{preLRS}(t[:i]) \neq \max[\text{preLRS}(t[:i])]_{t[:i-1]}^R$ 
10    Let  $k \leftarrow \min_{\prec} \{a, \max(\text{rex}(u) \cap \mathcal{N})\}$ ,  $v \leftarrow \text{Trans}(u, k - 1, 0)$ ,
     $\text{child}(u, Z(a, \text{len}(u))) \leftarrow \text{sink}$ , and  $u \leftarrow \text{SL}(u)$ ;
   //  $v$  corresponds to  $[\text{LRS}(t[:i])]_{t[:i-1]}^R$  and  $k = |\text{LRS}(t[:i])|$ 
11  if  $\text{len}(v) = k$  then Let  $\text{SL}(\text{sink}) \leftarrow v$ ; // No node split
12  else // Node split
13    Create a new node  $v'$ ; //  $v'$  corresponds to  $[\text{LRS}(t[:i])]_{t[:i]}^R$ 
14    Let  $\text{len}(v') \leftarrow k$ ;
    // In-coming edges of the new node
15    while  $\text{child}(u, Z(a, \text{len}(u))) = v$  do
16      Let  $\text{child}(u, Z(a, \text{len}(u))) \leftarrow v'$  and  $u \leftarrow \text{SL}(u)$ ;
    // Out-going edges of the new node
17    for each  $b \in \{b \in \text{rex}(v) \mid Z(b, k) \neq 0\}$  do
18      Let  $\text{child}(v', b) \leftarrow \text{child}(v, b)$ ;
19    if  $\text{Trans}(v, k, 0) \neq \text{Null}$  then Let  $\text{child}(v', 0) \leftarrow \text{Trans}(v, k, 0)$ ;
    // Suffix links
20    Let  $\text{SL}(v') \leftarrow \text{SL}(v)$ ,  $\text{SL}(v) \leftarrow v'$  and  $\text{SL}(\text{sink}) \leftarrow v'$ ;
21 return  $(V, E, \text{SL})$ ;
```

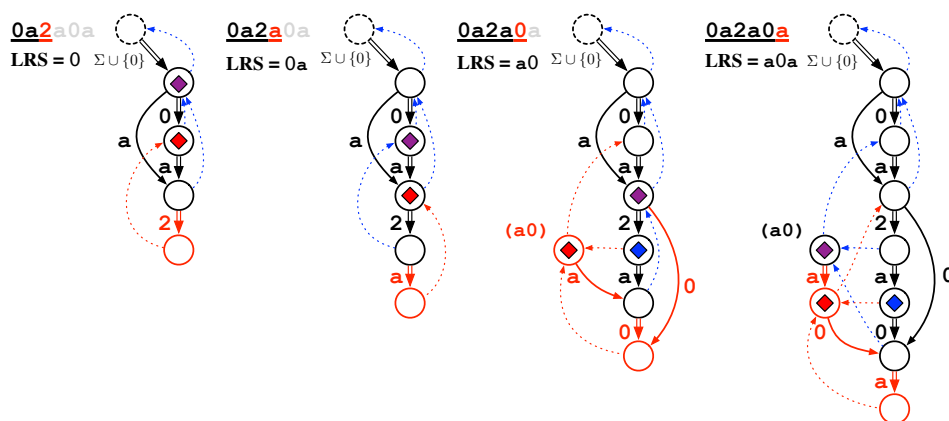
► **Lemma 12** (Suffix link update). *Suppose $V_{wa} = V_w \cup \{[wa]_{wa}^R\}$. Then, for each $u \in V_{wa}$,*

$$\text{SL}_{wa}(u) = \begin{cases} [\text{LRS}(wa)]_{wa}^R & \text{if } u = [wa]_{wa}^R, \\ \text{SL}_w(u) & \text{otherwise.} \end{cases}$$

Suppose $[x]_w^R \neq [x]_{wa}^R$ for $x = \text{LRS}(wa)$, i.e., $V_{wa} = V_w \setminus \{[y]_w^R\} \cup \{[wa]_{wa}^R, [x]_{wa}^R, [y]_{wa}^R\}$, where $y = \max[x]_w^R$. Then, for each $u \in V_{wa}$,

$$\text{SL}_{wa}(u) = \begin{cases} [x]_{wa}^R & \text{if } u \in \{[wa]_{wa}^R, [y]_{wa}^R\}, \\ \text{SL}_w([y]_w^R) & \text{if } u = [x]_{wa}^R, \\ \text{SL}_w(u) & \text{otherwise.} \end{cases}$$

Algorithm 3 constructs PDAWGs based on the above lemmas. An example of online construction of a PDAWG can be found in Figure 4. For technical convenience, like the standard DAWG construction algorithm, we add a dummy node \top to the PDAWG that has edges to the source node, denoted as ρ in Algorithm 3, labeled with all elements of



■ **Figure 4** A snapshot of left-to-right online construction of $\text{PDAWG}(T)$ with $T = \text{xaxaya}$ by Algorithm 3. Each figure shows $\text{PDAWG}(wa)$ for a prefix wa of $\text{prev}(T) = 0a2a0a$. Double arrows show primary edges. The new nodes, edges and suffix links are colored red. The purple, red and blue diamonds represent $[x']_{wa}^R$, $[x]_{wa}^R$ and $[y]_{wa}^R$, respectively, where $x' = \text{preLRS}(wa)$, $x = \text{LRS}(wa)$ and $y = \max[x]_{wa}^R$. When $x = \varepsilon$, the purple diamond is put on the dummy node \top .

$\Sigma \cup \{0\}$. This trick allows us to uniformly treat the special case where the LRS node is ρ , in which case the pre-LRS node is defined to be \top . In addition, we let $\text{SL}(\rho) = \top$. Each node u does not remember the elements of u but we remember $\text{len}(u) = |\max u|$. Note that $|\min u| = |\text{len}(\text{SL}(u))| + 1$. Hereafter we use functions SL , child , Trans , etc. without a subscript specifying a text, to refer to the data structure that the algorithm is manipulating, rather than the mathematical notion relative to the text. Of course, we design our algorithm so that those functions coincide with the corresponding mathematical notions.

Suppose we have constructed $\text{PDAWG}(w)$ and want to obtain $\text{PDAWG}(wa)$. The sink node of $\text{PDAWG}(w)$, denoted as *oldsink*, corresponds to $[w]_w^R$. We first make a new sink node *newsink* = $[wa]_{wa}^R$. Then we visit $u_i = \text{SL}^i(\text{oldsink})$ for $i = 1, 2, \dots, j$, until we find the pre-LRS node $u_j = [\text{preLRS}(wa)]_w^R$. By Lemma 8, we can identify u_j and $k' = |\text{preLRS}(wa)|$. For each node u_i with $i < j$, we make an edge labeled with $Z(a, \text{len}(u_i))$ pointing at *newsink* and, moreover, u_j also has an edge pointing at *newsink* if $k' < \text{len}(u_j)$ by Lemma 9. We then reach the LRS node $v = [\text{LRS}(wa)]_w^R = \text{Trans}(u_j, k', Z(a, k'))$. We compare $k = k' + 1 = |\text{LRS}(wa)|$ and $\text{len}(v)$ to decide whether the LRS node shall be split based on Lemma 6. If $|\text{LRS}(wa)| = \text{len}(v)$, the node v will not be split, in which case we obtain $\text{PDAWG}(wa)$ by making $\text{SL}(\text{newsink}) = v$ (Lemma 12).

Suppose $k < \text{len}(v)$. In this case, the LRS node v must be split. We reuse the old node v , which used to correspond to $[\text{LRS}(wa)]_w^R$, as a new node corresponding to $[\max[\text{LRS}(wa)]_w^R]_{wa}^R$, and create another new node v' for $[\text{LRS}(wa)]_{wa}^R$ with $\text{len}(v') = k$. Edges are determined in accordance with Lemmas 10 and 11. All in-coming edges from $\text{SL}^i([\text{preLRS}(wa)]_w^R)$ to v in $\text{PDAWG}(w)$ are redirected to v' , except when $\text{preLRS}(wa) \neq \max[\text{preLRS}(wa)]_w^R$ for $i = 0$. The out-going edges from v will be kept. We create out-going edges of v' referring to the corresponding transitions from v . If $(v, b, u) \in E$ with $Z(b, k) \neq 0$, then we add (v', b, u) to E . In addition, we add $(v', 0, \text{Trans}_w(v, 0, k))$ to E if $\text{Trans}(v, 0, k) \neq \text{Null}$. At last, suffix links among *newsink*, v , v' are determined in accordance with Lemma 12.

We conclude the subsection with the complexity of Algorithm 3. Let us call an edge (u, a, v) *primary* if $\max v = \max u \cdot a$, and *secondary* otherwise. The following lemma is an adaptation of the corresponding one for DAWGs by Blumer et al. [4].

► **Lemma 13.** *Let $\text{SC}_w(u) = \{\text{SL}_w^i(u) \mid i \geq 0\}$ for a node u . If $\text{PDAWG}(w)$ has a primary edge from u to v , then the total number of secondary edges from nodes in $\text{SC}_w(u)$ to nodes in $\text{SC}_w(v)$ is bounded by $|\text{SC}_w(u)| - |\text{SC}_w(v)| + |\Pi| + 1$.*

Proof. Let us count the number of edges from nodes in $\text{SC}_w(u)$ to $\text{SC}_w(v)$. Baker [2, Lemma 1] showed that in a parameterized suffix tree, each path from the root to a leaf has at most $|\Pi|$ nodes with *bad suffix links*. Through the duality of PDAWGs and parameterized suffix trees stated in Lemma 16, this means that $\text{SC}_w(v)$ contains at most $|\Pi| + 1$ nodes which has no in-coming primary edges, where the additional one node is the root of the PDAWG. Since each node has at most one in-coming primary edge, the number of primary edges in concern is at least $|\text{SC}_w(v)| - |\Pi| - 1$ in total. Since each node in $\text{SC}_w(u)$ has just one out-going edge to $\text{SC}_w(v)$, we obtain the lemma. ◀

► **Theorem 14.** *Given a string T of length n , Algorithm 3 constructs $\text{PDAWG}(T)$ in $O(n|\Pi| \log(|\Sigma| + |\Pi|))$ time and $O(n)$ space online, by reading T from left to right.*

Proof. Since the size of a PDAWG is bounded by $O(n)$ (Theorem 2) and nodes are monotonically added, it is enough to bound the number of edges and suffix links that are deleted. In each iteration of the **for** loop, at most one suffix link is deleted. So at most n suffix links are deleted in total. We count the number of edges whose target is altered from $v = [\text{LSR}(wa)]_w^R$ to $v' = [\text{LSR}(wa)]_{wa}^R$ on Line 15 when updating $\text{PDAWG}(w)$ to $\text{PDAWG}(wa)$. Let k_i be the number of such edges at the i -th iteration of the **for** loop. Note that those are all secondary edges from a node in $\text{SC}_w(u_0)$ for the pre-LRS node u_0 . By Lemma 13,

$$\begin{aligned} \sum_{i=1}^n k_i &\leq \sum_{i=1}^n (|\text{SC}_{wa}(w)| - |\text{SC}_{wa}(wa)| + |\Pi| + 1) \\ &\leq \sum_{i=1}^n (|\text{SC}_w(w)| - |\text{SC}_{wa}(wa)| + |\Pi| + 1) \\ &= |\text{SC}_\varepsilon(\varepsilon)| - |\text{SC}_t(t)| + (|\Pi| + 1)n \in O(|\Pi|n). \end{aligned}$$

Since the suffix links of $\text{PDAWG}(T)$ forms the p -suffix tree of \bar{T} (see Subsection 3.5), the following corollary is immediate from Theorem 14.

► **Corollary 15.** *The p -suffix tree of a string S of length n can be constructed in $O(n|\Pi| \log(|\Sigma| + |\Pi|))$ time and $O(n)$ space online, by reading S from right to left.*

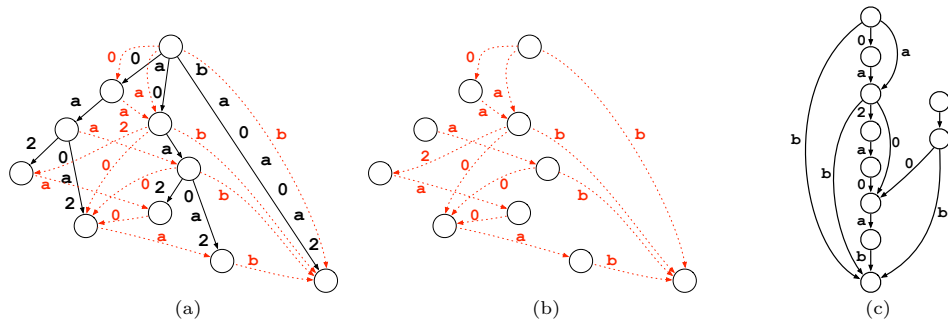
Differently from the online DAWG construction algorithm [4], we have the factor $|\Pi|$ in our algorithm complexity analysis. Actually our algorithm takes time proportional to the difference of the old and new PDAWGs modulo logarithmic factors, as long as the difference is defined so that the split node $[\text{LRS}(wa)]_w^R$ automatically becomes $[\max[\text{LRS}(wa)]_w^R]_{wa}^R$ rather than $[\text{LRS}(wa)]_{wa}^R$. In this sense, our algorithm is optimal. It is open whether we could improve the analysis.

3.5 Duality of PDAWGs and p -suffix trees

This subsection establishes the duality between parameterized suffix trees and PDAWGs. An example can be found in Figure 5. For this sake, we introduce the reverse of a pv -string and *Weiner links (reversed suffix links)* for parameterized suffix trees.

The “reverse” \tilde{x} of a pv -string x must satisfy that $\tilde{x} = \text{prev}(\bar{S})$ iff $x = \text{prev}(S)$ for any p -string $S \in (\Sigma \cup \Pi)^*$. For the empty string $\tilde{\varepsilon} = \varepsilon$. For $x \in (\Sigma \cup \mathcal{N})^*$ and $a \in \Sigma \cup \mathcal{N}$,

$$\tilde{xa} = \begin{cases} a\tilde{x} & \text{if } a \in \Sigma \cup \{0\}, \\ 0y & \text{otherwise,} \end{cases}$$



■ **Figure 5** (a) The parameterized suffix tree $\text{PSTree}(S)$ for $S = \text{baxayay}$ over $\Sigma = \{a, b\}$ and $\Pi = \{x, y\}$, augmented with the Weiner links (dashed red arcs). (b) The DAG consisting of the p-suffix tree nodes and the Weiner-links. (c) The PDAWG $\text{PDAWG}(T)$ for $T = \bar{S} = \text{yayaxab}$. The graphs (b) and (c) are isomorphic.

where y is obtained from \tilde{x} by replacing the a -th element by a , i.e., $y = \tilde{x}[: a - 1] \cdot a \cdot \tilde{x}[a + 1 :]$. This is well-defined if $x a$ is a pv-string. For example, for $T = \text{xaxy}$ with $a \in \Sigma$ and $x, y \in \Pi$, we have $\text{prev}(\bar{T}) = \overline{0a20} = 00a2 = \text{prev}(y x a x) = \text{prev}(\bar{T})$.

The parameterized suffix tree $\text{PSTree}(T)$ of a p-string T is the path-compacted (or Patricia) tree for $\text{PSuffix}(T)$. For any $z \in (\Sigma \cup \mathcal{N})^*$, For $\text{PSTree}(T)$, the Weiner links are defined as follows. Let v be a node in $\text{PSTree}(T)$ such that $v = \text{prev}(S)$ for some substring S of T , and $a \in \Sigma \cup \mathcal{N}$. Let $\alpha(a, v)$ be the pv-string such that

$$\alpha(a, v) = \begin{cases} av & \text{if } av \in \text{PFactor}(T) \text{ and } a \in \Sigma \cup \{0\}, \\ \text{prev}(S[a] \cdot S) & \text{if } \text{prev}(S[a] \cdot x) \in \text{PFactor}(T) \text{ and } a \in \mathcal{N} \setminus \{0\}, \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Then a Weiner link is a triple (v, a, u) such that $u = \alpha(a, v)y$, where $y \in (\Sigma \cup \mathcal{N})^*$ is the shortest string such that $\alpha(a, v)y$ is a node of $\text{PSTree}(T)$. The Weiner link (v, a, u) is said to be *explicit* if $u = \alpha(a, v)$, and *implicit* otherwise¹.

To establish the correspondence between $\text{PDAWG}(T)$ and $\text{PSTree}(\bar{T})$ easily, here we rename the nodes $[x]_w^R$ of $\text{PDAWG}(T)$ to be $\max[x]_w^R$ where $w = \text{prev}(T)$.

► **Theorem 16.** *The following correspondence between $\text{PDAWG}(T) = (V_D, E_D)$ and $\text{PSTree}(\bar{T}) = (V_T, E_T)$ holds.*

- (1) $\text{PDAWG}(T)$ has a node $x \in V_D$ iff $\text{PSTree}(\bar{T})$ has a node $\tilde{x} \in V_T$.
- (2) $\text{PDAWG}(T)$ has a primary edge $(x, a, y) \in E_D$ iff $\text{PSTree}(\bar{T})$ has an explicit Weiner link $(\tilde{x}, a, \tilde{y})$.
- (3) $\text{PDAWG}(T)$ has a secondary edge $(x, a, y) \in E_D$ iff $\text{PSTree}(\bar{T})$ has an implicit Weiner link $(\tilde{x}, a, \tilde{y})$.
- (4) $\text{PDAWG}(T)$ has a suffix link from $\tilde{x}y$ to \tilde{x} iff $\text{PSTree}(\bar{T})$ has an edge $(x, y, xy) \in E_T$.

¹ Explicit Weiner links are essentially the same as the reversed suffix links used for right-to-left online construction of parameterized position heaps [8].

References

- 1 Brenda S. Baker. A theory of parameterized pattern matching: algorithms and applications. In *STOC 1993*, pages 71–80, 1993.
- 2 Brenda S. Baker. Parameterized pattern matching: Algorithms and applications. *Journal of Computer and System Sciences*, 52(1):28–42, 1996.
- 3 Richard Beal and Donald A. Adjeroh. p-suffix sorting as arithmetic coding. *J. Discrete Algorithms*, 16:151–169, 2012.
- 4 Anselm Blumer, Janet Blumer, David Haussler, Andrzej Ehrenfeucht, Mu-Tian Chen, and Joel Seiferas. The smallest automation recognizing the subwords of a text. *Theoretical computer science*, 40:31–55, 1985.
- 5 Maxime Crochemore. Transducers and repetitions. *Theor. Comput. Sci.*, 45(1):63–86, 1986.
- 6 Satoshi Deguchi, Fumihito Higashijima, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda. Parameterized suffix arrays for binary strings. In *PSC 2008*, pages 84–94, 2008.
- 7 Diptarama, Takashi Katsura, Yuhei Otomo, Kazuyuki Narisawa, and Ayumi Shinohara. Position heaps for parameterized strings. In *CPM 2017*, pages 8:1–8:13, 2017. doi:10.4230/LIPIcs.CPM.2017.8.
- 8 Noriki Fujisato, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Right-to-left online construction of parameterized position heaps. In *PSC 2018*, pages 91–102, 2018.
- 9 Noriki Fujisato, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. Direct Linear Time Construction of Parameterized Suffix and LCP Arrays for Constant Alphabets. In *SPIRE 2019*, pages 382–391. Springer International Publishing, 2019.
- 10 Noriki Fujisato, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. The Parameterized Position Heap of a Trie. In *CIAC 2019*, pages 237–248. Springer International Publishing, 2019.
- 11 Arnab Ganguly, Rahul Shah, and Sharma V. Thankachan. pBWT: Achieving succinct data structures for parameterized pattern matching and related problems. In *SODA 2017*, pages 397–407, 2017.
- 12 Tomohiro I, Satoshi Deguchi, Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda. Light-weight parameterized suffix array construction. In *IWOCA 2009*, pages 312–323, 2009.
- 13 S. Rao Kosaraju. Faster algorithms for the construction of parameterized suffix trees (preliminary version). In *FOCS 1995*, pages 631–637, 1995.
- 14 Taehyung Lee, Joong Chae Na, and Kunsoo Park. On-line construction of parameterized suffix trees for large alphabets. *Inf. Process. Lett.*, 111(5):201–207, 2011.
- 15 Juan Mendivelso and Yoan Pinzón. Parameterized matching: Solutions and extensions. In *Proc. PSC 2015*, pages 118–131, 2015.
- 16 Juan Mendivelso, Sharma V. Thankachan, and Yoan Pinzón. A brief history of parameterized matching problems. *Discrete Applied Mathematics*, 2018. Available online. doi:10.1016/j.dam.2018.07.017.
- 17 Katsuhito Nakashima, Noriki Fujisato, Diptarama Hendrian, Yuto Nakashima, Ryo Yoshinaka, Shunsuke Inenaga, Hideo Bannai, Ayumi Shinohara, and Masayuki Takeda. DAWGs for parameterized matching: online construction and related indexing structures. *CoRR*, abs/2002.06786, 2020. URL: <https://arxiv.org/abs/2002.06786>.
- 18 Tetsuo Shibuya. Generalization of a suffix tree for RNA structural pattern matching. *Algorithmica*, 39(1):1–19, 2004.