The Average Profile of Suffix Trees

Mark Daniel Ward *
Department of Mathematics
University of Pennsylvania
Philadelphia, PA 19104–6395
ward2@math.upenn.edu

Abstract

The internal profile of a tree structure denotes the number of internal nodes found at a specific level of the tree. Similarly, the external profile denotes the number of leaves on a level. The profile is of great interest because of its intimate connection to many other parameters of trees. For instance, the depth, fill-up level, height, path length, shortest path, and size of trees can each be interpreted in terms of the profile.

The current study is motivated by the work of Park et al. [22], which was a comprehensive study of the profile of tries constructed from *independent* strings (also, each string generated by a memoryless source). In the present paper, however, we consider suffix trees, which are constructed from suffixes of a common string. The dependency between suffixes demands a careful, intricate treatment of overlaps in words.

We precisely analyze the average internal and external profiles of suffix trees generated by a memoryless source. We utilize combinatorics on words (in particular, autocorrelation, i.e., the degree to which a word overlaps with itself) generating functions, singularity analysis, and the Mellin transform. We make comparisons of the average profile of suffix trees to the average profile of tries constructed from independent strings. We emphasize that our methods are extensible to higher moments. The present report describes the first moment of both the internal and external profiles of suffix trees.

1 Introduction.

The profile parameter of a tree data structure concerns the number of nodes located at a certain level of the tree. In other words, the profile of a tree is the enumeration of nodes at a given distance from the root node, i.e., at a given depth in the tree. In particular, the internal profile of a tree denotes the number of internal (i.e., branching) nodes located on a given level. The external profile of a tree is the number of external (i.e., leaf) nodes located on a given level.

The profile parameter of various tree data structures has recently garnered a great deal of attention in the literature. One justification for the extensive attempt to understand the profile parameter is its relevance to a great number of other tree parameters, for instance, the depth, fill-up level, height, path length,

shortest path, and size of trees can each be interpreted in terms of the profile. Also, for instance, the internal profile at level k of a suffix tree is exactly the number of words of length k in the tree which have two or more occurrences. This interpretation becomes useful in applications in which repeated occurrences of words are crucial, for instance, in data compression, leader elections, molecular biology, etc.

Due to space constraints, we refer the interested reader to the recent seminal paper of Park et al. [22] (which inherits from [20], [21], and [23]). Their 60 page paper is a comprehensive study of the internal and external profiles of tries constructed from *independent* strings (each string is also generated by a memoryless source). In contrast, in the present report, we consider suffix trees, which are constructed from suffixes of a common string. We emphasize that the dependency between suffixes repeatedly requires us to treat the overlap exhibited by a word with itself.

We discuss the average internal and external profiles of suffix trees. In this report, each suffix tree is built from a memoryless source; nonetheless, of course, the suffixes are highly dependent on each other. For this reason, we must consider the extent to which words overlap with themselves. For this purpose, we utilize recent results from the literature of combinatorics on words. Other techniques that we use include generating functions, singularity analysis, and the Mellin transform.

We emphasize the difficulty in the analysis of the profile even in tries built from independent strings (again, we refer to the lengthy report [22]). Thus, we content ourselves in the present report with proving that the profiles in tries (built from independent strings) and in suffix trees asymptotically have the same behavior for a wide range of the relevant parameters. Since the behavior of the profiles in tries constructed from independent strings was described so thoroughly in [22], we content ourselves here with making intricate comparisons to the asymptotic behavior of the analogous profiles in

suffix trees.

We must consider the relationship between the number n of strings inserted into a suffix tree and the level k of the tree for which we extract the profile. In [22], the behavior of the the profile is categorized according to the relationship between n and k. For this reason, we tried to derive the most general type of result about profiles in suffix trees as compared to profiles in tries constructed from independent strings. Theorems 3.1 and 3.2 are valid for all k and n. Therefore, whether k is a function of n, or (on the other hand) if k is treated independently of n, the theorems below still hold. The ϵ and μ in Theorems 3.1 and 3.2 do *not* depend on n or k; i.e., the theorems are true for all n and k.

The higher moments of the profile of suffix trees— as well as the exact distribution—are significantly more difficult with the methods applied in this report. For instance, the second moment of the profile of suffix trees requires not only a careful account of the degree to which a word overlaps with itself (i.e., the autocorrelation of a word), but also the degree to which two distinct words of the same length overlap with each other (i.e., the correlation of two distinct words). Higher moments appear to require an even more extensive application of correlation amongst sets of words. For this reason, we present here only the average behavior of the internal and external profiles of suffix trees. Although we have looked extensively at the second moment of the profile of suffix trees, we are limited by the space constraints of this report. We expect to present a detailed analysis of the second moment of the profile of suffix trees in the near future, in a separate report.

We mention only a sampling of relevant papers: For some results about the profile of other types of trees, we suggest [2], [3], [4], [5], [6], [10], and [13]. For additional material on suffix trees, we recommend [1], [7], [8], [12], [14], [16], [15], and [27]. For various applications of combinatorics on words, we mention [11], [14], [15], [24], and [25]. Of course, a variety of other excellent reports on suffix trees and combinatorics on words appear in the literature.

In the present paper, we first establish some definitions, and then we present the main results. Basically, we report that the average profiles of suffix trees and of tries constructed from independent strings asymptotically have the same behaviors. After stating the main results, the remainder of the paper is dedicated to the proofs. The proofs include a discussion of combinatorics on words; generating functions for the average internal and external profiles in suffix trees and in tries constructed from independent strings; singularity analysis; residue extraction; and finally an application of the Mellin transform.

2 Definitions.

Throughout the discussion below, we work with binary words, i.e., words from $\{0,1\}^* = \mathcal{A}^*$. We first briefly describe the means of constructing a trie data structure. Consider a set of n binary strings $X^{(1)}, \ldots, X^{(n)}$, where the ith such string is denoted by

$$X^{(i)} = X_1^{(i)} X_2^{(i)} X_3^{(i)} \ldots .$$

We insert the strings into a trie, starting at the root node, in a recursive manner: If $n = 0$, the trie is empty. If $n = 1$, the trie is a leaf. If $n \geq 2$, then at the ith stage in the construction, we partition the current set of words into two sets, corresponding to whether the ith letter of each word is a 0 or 1. All words under consideration with ith letter 0 (respectively, 1) are placed into the left (respectively, right) subtree.

In the case where the n strings $X^{(1)}, \ldots, X^{(n)}$ are chosen independently of each other, we use a "hat" symbol on the relevant parameters. In a trie constructed from independent strings, we use the following notation for consistency with Park et al. [22]. The external profile $\widehat{B}_{n,k}$ denotes the number of leaf nodes found on the kth level of the trie (in other words, the number of leaf nodes at a distance k from the root). Similarly, the internal profile $\widehat{U}_{n,k}$ denotes the number of internal (i.e., branching) nodes found at the k level of the trie.

Now we consider a different model, in which the strings $X^{(1)}, \ldots, X^{(n)}$ are dependent on each other. We let $X = X_1 X_2 X_3 \ldots$ denote a binary string, where the X_i's are chosen according to a probability model to be specified shortly. We write $X^{(i)} = X_i X_{i+1} X_{i+2} \ldots$ to denote the ith suffix of X. The trie that is constructed from the first n suffixes of X, namely from $X^{(1)}, \ldots, X^{(n)}$, is called a suffix tree. The external profile $B_{n,k}$ in such a suffix tree denotes the number of external nodes (i.e., leaf nodes) found at distance k from the root. The internal profile $U_{n,k}$ denotes the analogous number of internal (i.e., branching) nodes.

Throughout the remainder of the paper, we focus attention on $X^{(i)}$ and X such that the various $X_j^{(i)}$ and X_j's are independent from each other and are each generated by a stationary, Bernoulli source. Thus, we write $\mathbf{P}(w)$ for the stationary probability of the occurrence of a word w; namely, $\mathbf{P}(w) = p^m q^n$ if w contains exactly m "0"s and n "1"s. We have $\mathbf{P}(X_j = 0) = p$ and $\mathbf{P}(X_j = 1) = q$ in the suffix tree model; similarly $\mathbf{P}(X_j^{(i)} = 0) = p$ and $\mathbf{P}(X_j^{(i)} = 1) = q$ in the trie model constructed from independent strings. Without loss of generality, throughout the discussion we assume that $0 < q \leq p < 1$. We define $\delta = \sqrt{p}$. We choose $c > 0$ such that $q^{-c}\delta < 1$, and we choose ϵ with $0 < \epsilon < c$. Finally, we define $\mu = q^{-c}\delta$ for ease of

notation.

At various points in the paper we use words $w \in \mathcal{A}^*$, and $\alpha, \beta \in \mathcal{A}$ with $\alpha = 1 - \beta$. In other words, the ordered pair (α, β) denotes either the pair $(0, 1)$ or the pair $(1, 0)$. This allows us to consider both $w\alpha$ and its sibling $w\beta$.

3 Main Results.

We point out once again that [22] presents a plethora of results concerning the average behavior of the internal and external profiles of tries constructed from independent strings. In particular, [22] classifies the behavior of such profiles according to the relationship between n and k. Since [22] is such a comprehensive description of the profile in tries constructed from independent strings, it is very fruitful to make a comparison of such profiles to the profiles of suffix trees. Many of the results about the average profile in tries constructed from independent strings can be translated to analogous results concerning the average profiles of suffix trees.

For this reason, we tried to derive the most general type of result about the relationship between the profiles in suffix trees as compared to profiles in tries constructed from independent strings. Theorems 3.1 and 3.2 are valid for all relationships between k and n. Therefore, whether k is a function of n, or (on the other hand) if k is treated independently of n, the theorems below still hold. Also, we emphasize that the ϵ and μ in Theorems 3.1 and 3.2 below do *not* depend on n or k; in other words, these are valid comparisons for all n and k.

THEOREM 3.1. *Recall* $0 < q \leq p < 1$, *and also* $\delta = \sqrt{p}$. *Consider* $c > 0$ *such that* $\mu := q^{-c}\delta < 1$, *and* ϵ *with* $0 < \epsilon < c$.

The difference in the average internal profile $U_{n,k}$ *of a suffix tree (at level k, using the first n suffixes of a common word) versus the average internal profile* $\widehat{U}_{n,k}$ *of a trie (constructed from n independent strings) is asymptotically negligible. The difference satisfies*

$$\mathbb{E}(U_{n,k}) - \mathbb{E}(\widehat{U}_{n,k}) = O(n^{-\epsilon}\mu^k).$$

THEOREM 3.2. *Recall* $0 < q \leq p < 1$, *and also* $\delta = \sqrt{p}$. *Consider* $c > 0$ *such that* $\mu := q^{-c}\delta < 1$, *and* ϵ *with* $0 < \epsilon < c$.

The difference in the average external profile $B_{n,k}$ *of a suffix tree (at level k, using the first n suffixes of a common word) versus the average internal profile* $\widehat{B}_{n,k}$ *of a trie (constructed from n independent strings) is asymptotically negligible. The difference satisfies*

$$\mathbb{E}(B_{n,k}) - \mathbb{E}(\widehat{B}_{n,k}) = O(n^{-\epsilon}\mu^k).$$

We note that, in tries constructed from independent strings, the nodes at the kth level appear for $k \sim \log n$. (In other words, the case $k \sim \log n$ is, in many respects, the most interesting case.) So we state immediate corollaries of Theorems 3.1 and 3.2 in the analogous case, namely, when $k = a \log n$, for $a > 0$ constant.

COROLLARY 3.1. *Recall* $0 < q \leq p < 1$, *and also* $\delta = \sqrt{p}$. *Consider* $c > 0$ *such that* $\mu := q^{-c}\delta < 1$, *and* ϵ *with* $0 < \epsilon < c$.

Let $k = a \log n$, *where* $a > 0$ *is constant. The difference in the average internal profile* $U_{n,k}$ *of a suffix tree (at level k, using the first n suffixes of a common word) versus the average internal profile* $\widehat{U}_{n,k}$ *of a trie (constructed from n independent strings) is asymptotically negligible. The difference satisfies*

$$\mathbb{E}(U_{n,k}) - \mathbb{E}(\widehat{U}_{n,k}) = O(n^{-\delta}),$$

where $\delta = \epsilon + c \log(1/\mu) > 0$.

COROLLARY 3.2. *Recall* $0 < q \leq p < 1$, *and also* $\delta = \sqrt{p}$. *Consider* $c > 0$ *such that* $\mu := q^{-c}\delta < 1$, *and* ϵ *with* $0 < \epsilon < c$.

Let $k = a \log n$, *where* $a > 0$ *is constant. The difference in the average external profile* $B_{n,k}$ *of a suffix tree (at level k, using the first n suffixes of a common word) versus the average internal profile* $\widehat{B}_{n,k}$ *of a trie (constructed from n independent strings) is asymptotically negligible. The difference satisfies*

$$\mathbb{E}(B_{n,k}) - \mathbb{E}(\widehat{B}_{n,k}) = O(n^{-\delta}),$$

where $\delta = \epsilon + c \log(1/\mu) > 0$.

The methodology used to prove these theorems about the average profiles appears to be fully extensible to similar proofs concerning the second moment of the profiles.

We note that Theorems 3.1 and 3.2 are quite similar in nature. So it is not surprising that the methodology for both proofs is the same. As we will see, however, the proof of Theorem 3.1 is slightly more manageable at the point in which we begin to extract residues. For this reason, due to space constraints, we give more details about the proof of Theorem 3.1. We assure the reader that analogous techniques can be utilized at the end of the proof of Theorem 3.2.

No proof of Corollary 3.1 or 3.2 is necessary; we simply have substituted $k = a \log n$ into the Theorems 3.1 and 3.2.

The remainder of the paper is dedicated to establishing these two theorems.

4 Combinatorics on Words.

We utilize some results from the literature of combinatorics on words. For a starting point to the theory of combinatorics on words, we refer the reader to [11], [14], [15], [24], and [25]; this is merely a sampling of the growing literature in this area. For a collection of recent results, see the three volumes edited by Lothaire [17], [18], and [19].

At the heart of the theory of combinatorics on words is a precise characterization of the extent to which a word overlaps with itself. For this purpose, for each word w of length m (i.e., $w \in \mathcal{A}^m$), we define the autocorrelation polynomial of w as

$$(4.1) \qquad S_w(z) = \sum_{i \in \mathcal{P}(w)} \mathbf{P}(w_{i+1} \ldots w_m) z^{m-i} \,,$$

where $\mathcal{P}(w)$ denotes the set of i's satisfying $w_1 \ldots w_i = w_{m-i+1} \ldots w_m$. In other words, for each $i \in \mathcal{P}(w)$, the prefix of w of length i is identical to the suffix of w of length i.

Now we define some useful languages—and their associated generating functions—frequently used in combinatorics on words. We write

$$\mathcal{R}_w = \{v \in \mathcal{A}^* \mid v \text{ contains exactly one occurrence}$$
$$\text{of } w, \text{ located at the right end}\} \,,$$
$$\mathcal{M}_w = \{v \in \mathcal{A}^* \mid wv \text{ contains exactly two occurrences}$$
$$\text{of } w, \text{ located at the left and right ends}\} \,,$$
$$\mathcal{U}_w = \{v \in \mathcal{A}^* \mid wv \text{ contains exactly one occurrence}$$
$$\text{of } w, \text{ located at the left end}\} \,.$$

We write the generating functions associated with these languages as

$$R_w(z) = \sum_{v \in \mathcal{R}_w} \mathbf{P}(v) z^{|v|} \,,$$
$$M_w(z) = \sum_{v \in \mathcal{M}_w} \mathbf{P}(v|w) z^{|v|} \,,$$
$$U_w(z) = \sum_{v \in \mathcal{U}_w} \mathbf{P}(v|w) z^{|v|} \,,$$

where $\mathbf{P}(v|w) := \mathbf{P}(wv)/\mathbf{P}(w)$. It is well-known (see, for instance, [15]) that these generating functions can easily be expressed in terms of $S_w(z)$. If $w \in \mathcal{A}^m$, and if we define

$$D_w(z) = (1-z)S_w(z) + \mathbf{P}(w)z^m \,,$$

then we have

$$R_w(z) = \frac{\mathbf{P}(w)z^m}{D_w(z)} \,,$$
$$M_w(z) = 1 + \frac{z-1}{D_w(z)} \,,$$
$$U_w(z) = \frac{1}{D_w(z)} \,.$$

We also need to utilize a more general theory of counting occurrences of words found in a set of restricted words. In other words, we need the ability to simultaneously count the number of occurrences of two words, say $H_1, H_2 \in \mathcal{A}^*$, within a longer word. Of course, in this more general situation, the overlaps of H_1 and H_2 can occur frequently or rarely, depending on the forms of H_1 and H_2. For each pair H_1, H_2 we are interested in counting, we must contend with overlaps of H_1 with itself, overlaps of H_2 with itself, and also the overlaps of H_1 and H_2 with each other. This generalized theory of correlations between pairs of words is discussed, for instance, in [15] and [24]. In this report, we restrict attention to counting the number of occurrences of two related words, of the form $w\alpha$ and $w\beta$, for $w \in \mathcal{A}^*$ and $\alpha = 1 - \beta$ (in other words, $\{\alpha, \beta\} = \mathcal{A}$). For this reason, we do not need the full generality of [15] and [24]. When we restrict attention to the number of occurrences of words of the form $w\alpha$ and $w\beta$, the theory becomes somewhat simpler. In this specific case, the languages and generating functions that we need can be reduced to very applicable forms.

The following are two useful languages, along with their generating functions, that we will use in the proofs of Theorems 3.1 and 3.2. We write

$$\widetilde{\mathcal{R}}_{w\alpha} = \{v \in \mathcal{A}^* \mid v \text{ contains exactly one occurrence}$$
$$\text{of } w\alpha, \text{ located at the right end,}$$
$$\text{and } no \text{ occurrences of } w\beta\} \,,$$
$$\widetilde{\mathcal{U}}_{w\alpha} = \{v \in \mathcal{A}^* \mid w\alpha v \text{ contains exactly one occurrence}$$
$$\text{of } w\alpha, \text{ located at the left end,}$$
$$\text{and } no \text{ occurrences of } w\beta\} \,.$$

The associated generating functions are

$$\widetilde{R}_{w\alpha}(z) = \sum_{v \in \widetilde{\mathcal{R}}_{w\alpha}} \mathbf{P}(v) z^{|v|} \,,$$
$$\widetilde{U}_{w\alpha}(z) = \sum_{v \in \widetilde{\mathcal{U}}_{w\alpha}} \mathbf{P}(v|w\alpha) z^{|v|} \,.$$

The languages $\widetilde{\mathcal{R}}_{w\beta}$ and $\widetilde{\mathcal{U}}_{w\beta}$, and their associated generating functions $\widetilde{R}_{w\beta}(z)$ and $\widetilde{U}_{w\beta}(z)$, are defined in an analogous way. A tilde over a language or its generating function is intended to denote the consideration

of both $w\alpha$ and $w\beta$. (Contrast this with the single-variable equivalents defined earlier, which are used for enumerating occurrences of a single word.)

In order to determine the generating functions of $\widetilde{\mathcal{R}}_{w\alpha}$, $\widetilde{\mathcal{R}}_{w\beta}$, $\widetilde{\mathcal{U}}_{w\alpha}$, and $\widetilde{\mathcal{U}}_{w\beta}$, we need to define a few languages, generating functions, and matrices from the generalized theory of combinatorics on words (see [15] and [24]). Again, we emphasize that we are interested in enumerating the number of (possibly overlapping) occurrences of the words $H_1 = w\alpha$ and $H_2 = w\beta$ within longer words.

First, we define

$$\mathcal{A}_{w\alpha,w\beta} = \{v \in \mathcal{A}^* \mid |v| \leq |w| \text{ and } w\alpha v \text{ has an}$$
$$\text{occurrence of } w\beta \text{ at the right end}\},$$

and, in an analogous way, we define

$$\mathcal{A}_{w\alpha,w\alpha} = \{v \in \mathcal{A}^* \mid |v| \leq |w| \text{ and } w\alpha v \text{ has an}$$
$$\text{occurrence of } w\alpha \text{ at the right end}\},$$
$$\mathcal{A}_{w\beta,w\alpha} = \{v \in \mathcal{A}^* \mid |v| \leq |w| \text{ and } w\beta v \text{ has an}$$
$$\text{occurrence of } w\alpha \text{ at the right end}\},$$
$$\mathcal{A}_{w\beta,w\beta} = \{v \in \mathcal{A}^* \mid |v| \leq |w| \text{ and } w\beta v \text{ has an}$$
$$\text{occurrence of } w\beta \text{ at the right end}\}.$$

Then

$$A_{w\alpha,w\beta}(z) = \sum_{v \in \mathcal{A}_{w\alpha,w\beta}} \mathbf{P}(v)z^{|v|}$$

is the associated correlation polynomial of $(w\alpha, w\beta)$. The autocorrelation polynomials $A_{w\alpha,w\alpha}(z)$, $A_{w\beta,w\alpha}(z)$, and $A_{w\beta,w\beta}(z)$ are defined in an analogous way.

We emphasize that $A_{w\alpha,w\alpha}(z)$ is exactly the autocorrelation polynomial $S_{w\alpha}(z)$.

Next, we define the correlation matrix

$$\mathbb{A}_w(z) = \begin{bmatrix} A_{w\alpha,w\alpha}(z) & A_{w\alpha,w\beta}(z) \\ A_{w\beta,w\alpha}(z) & A_{w\beta,w\beta}(z) \end{bmatrix}.$$

We observe

$$\mathbb{A}_w(z) = \begin{bmatrix} S_{w\alpha}(z) & (S_{w\alpha}(z)-1)\frac{\mathbf{P}(\beta)}{\mathbf{P}(\alpha)} \\ (S_{w\beta}(z)-1)\frac{\mathbf{P}(\alpha)}{\mathbf{P}(\beta)} & S_{w\beta}(z) \end{bmatrix}.$$

We also define the following matrix for $w \in \mathcal{A}^m$,

$$\mathbb{D}_w(z) = (1-z)\mathbb{A}(z) + \begin{bmatrix} \mathbf{P}(\alpha) & \mathbf{P}(\beta) \\ \mathbf{P}(\alpha) & \mathbf{P}(\beta) \end{bmatrix} \mathbf{P}(w)z^{m+1},$$

which will aid us in finding useful formulas for $\widetilde{R}_{w\alpha}(z)$,

$\widetilde{U}_{w\alpha}(z)$, $\widetilde{R}_{w\beta}(z)$, and $\widetilde{U}_{w\beta}(z)$. We easily compute

$$\mathbb{D}_w(z)^{-1} = \frac{1}{\det(\mathbb{D}_w(z))} \Bigg($$
$$(1-z)\begin{bmatrix} S_{w\beta}(z) & (1-S_{w\alpha}(z))\frac{\mathbf{P}(\beta)}{\mathbf{P}(\alpha)} \\ (1-S_{w\beta}(z))\frac{\mathbf{P}(\alpha)}{\mathbf{P}(\beta)} & S_{w\alpha}(z) \end{bmatrix}$$
$$+ \begin{bmatrix} \mathbf{P}(\beta) & -\mathbf{P}(\beta) \\ -\mathbf{P}(\alpha) & \mathbf{P}(\alpha) \end{bmatrix} \mathbf{P}(w)z^{m+1} \Bigg)$$

for $w \in \mathcal{A}^m$. Also

$$\det(\mathbb{D}_w(z)) = (1-z)D_w(z).$$

From the generalized theory of combinatorics on words, as in [15] and [24], we know that, for $w \in \mathcal{A}^m$,

$$(\widetilde{R}_{w\alpha}(z), \widetilde{R}_{w\beta}(z)) = (\mathbf{P}(w\alpha)z^{m+1}, \mathbf{P}(w\beta)z^{m+1})\mathbb{D}(z)^{-1}$$

and

$$\begin{bmatrix} \widetilde{U}_{w\alpha}(z) \\ \widetilde{U}_{w\beta}(z) \end{bmatrix} = \mathbb{D}(z)^{-1}\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

These formulas are easily simplified to reveal

$$(\widetilde{R}_{w\alpha}(z), \widetilde{R}_{w\beta}(z)) = \frac{(\mathbf{P}(w\alpha)z^{m+1}, \ \mathbf{P}(w\beta)z^{m+1})}{D_w(z)}$$

and

$$\begin{bmatrix} \widetilde{U}_{w\alpha}(z) \\ \widetilde{U}_{w\beta}(z) \end{bmatrix} = \frac{1}{D_w(z)}\begin{bmatrix} S_{w\beta}(z) + (1-S_{w\alpha}(z))\frac{\mathbf{P}(\beta)}{\mathbf{P}(\alpha)} \\ (1-S_{w\beta}(z))\frac{\mathbf{P}(\alpha)}{\mathbf{P}(\beta)} + S_{w\alpha}(z) \end{bmatrix}.$$

The generating functions $\widetilde{R}_{w\alpha}(z)$, $\widetilde{U}_{w\alpha}(z)$, $\widetilde{R}_{w\beta}(z)$, and $\widetilde{U}_{w\beta}(z)$ allow us to measure occurrences of words with overlaps. These generating functions and their associated languages are crucial in some of the arguments below.

5 Generating Functions.

Now we describe the generating functions for the average internal and external profiles of suffix trees and also for tries constructed from independent strings. Analogous, but more complicated, generating functions for the second moments of the profiles can be established using a similar methodology, but more intricate string comparisons must be utilized. So we save derivations for the second moments for a subsequent paper to appear in the near future.

We first briefly recall the models under consideration, namely, suffix trees and the tries constructed from independent strings. We have $X = X_1X_2X_3\ldots$, which is a binary string, where the X_j's are chosen

according to an (independent) Bernoulli source, with $\mathbf{P}(X_j = 0) = p$ and $\mathbf{P}(X_j = 1) = q$. The first n suffixes of X are used to generate a suffix tree. Also, we construct a trie from a set of n independent, binary strings $X^{(1)}, \ldots, X^{(n)}$, where the ith such string is denoted by $X^{(i)} = X_1^{(i)} X_2^{(i)} X_3^{(i)} \ldots$, with $\mathbf{P}(X_j^{(i)} = 0) = p$ and $\mathbf{P}(X_j^{(i)} = 1) = q$.

For ease of notation, we write $U(z) = \sum_{n \geq 0} \mathbb{E}(U_{n,k}) z^n$ and $B(z) = \sum_{n \geq 0} \mathbb{E}(B_{n,k}) z^n$ as the ordinary generating functions for the internal and external profiles (respectively) of a suffix tree. Similarly, we define $\widehat{U}(z) = \sum_{n \geq 0} \mathbb{E}(\widehat{U}_{n,k}) z^n$ and $\widehat{B}(z) = \sum_{n \geq 0} \mathbb{E}(\widehat{B}_{n,k}) z^n$ as the analogous OGFs for the profiles of a trie constructed from independent strings.

5.1 Generating Function for the Internal Profile of a Suffix Tree.
We use combinatorics on words to describe the internal profile at level k of a suffix tree. Thus, we restrict our attention to the prefixes of length k of the first n suffixes of X.

OBSERVATION 1. *The internal profile $U_{n,k}$ at level k of a suffix tree constructed from the first n suffixes of X is exactly the sum of the number of words $w \in \mathcal{A}^k$ with the property that $w = X_i \ldots X_{i+k-1}$ for at least two values of i with $1 \leq i \leq n$. In other words, $U_{n,k}$ is the sum of the number of words $w \in \mathcal{A}^k$ that appear as prefixes of at least two of first n suffixes of X.*

LEMMA 5.1. *Consider the polynomial $D_w(z) = (1-z) S_w(z) + \mathbf{P}(w) z^m$ associated with a word $w \in \mathcal{A}^m$, where $S_w(z) = \sum_{i \in \mathcal{P}(w)} \mathbf{P}(w_{i+1} \ldots w_m) z^{m-i}$ denotes the autocorrelation polynomial of w.*

The generating function for the average internal profile at level k in a suffix tree is

$$U(z) = \sum_{w \in \mathcal{A}^k} \frac{\mathbf{P}(w) z (D_w(z) - (1-z))}{(1-z) D_w(z)^2}.$$

Proof. We observe that $w \in \mathcal{A}^k$ makes a contribution to the internal profile at level k if and only if X begins with a word from $R_w M_w \mathcal{A}^*$ of length $n + k - 1$, which happens with probability

$$[z^{n+k-1}] \left(\frac{R_w(z) M_w(z)}{1-z} \right).$$

It follows immediately that

$$\sum_{n \geq 0} \mathbb{E}(U_{n,k}) z^n = \sum_{w \in \mathcal{A}^k} \frac{R_w(z) M_w(z)}{(1-z) z^{k-1}}.$$

Since (see, for instance, [14], [15], [24], and [25]) we have $R_w(z) M_w(z) = \mathbf{P}(w) z^k (D_w(z) - (1-z))/D_w(z)^2$, the lemma follows immediately. ∎

5.2 Generating Function for the External Profile of a Suffix Tree.
We again use combinatorics on words, this time to determine the external profile at level k of a suffix tree.

We are concerned with level k of the suffix tree, so we let $\mathcal{P}_n$ denote the set of prefixes of length k of the suffixes of X. We consider words of the form $w\alpha$, where $w \in \mathcal{A}^{k-1}$ and $\alpha \in \mathcal{A}$. We let $\beta = 1 - \alpha$, so that (α, β) is either the pair $(0,1)$ or $(1,0)$. In terms of trees, $w\alpha$ and $w\beta$ can naturally be viewed as siblings, i.e., children of the same parent. For example, if $w\alpha = 10010$ then $w\beta = 10011$.

OBSERVATION 2. *The external profile $B_{n,k}$ at level k of a suffix tree constructed from the first n suffixes of X is exactly the sum of the number of words $w\alpha$ for which two conditions both hold: (1.) $w\alpha$ appears in $\mathcal{P}_n$ exactly once, and (2.) $w\beta$ appears in $\mathcal{P}_n$ at least once. In other words, $B_{n,k}$ is the sum of the number of words $w\alpha$ such that $w\alpha = X_i \ldots X_{i+k-1}$ for exactly one value of i with $1 \leq i \leq n$, while $w\beta = X_j \ldots X_{j+k-1}$ for at least one value of j with $1 \leq j \leq n$.*

LEMMA 5.2. *Consider the polynomial $D_w(z) = (1-z) S_w(z) + \mathbf{P}(w) z^m$ associated with a word $w \in \mathcal{A}^m$, where $S_w(z) = \sum_{i \in \mathcal{P}(w)} \mathbf{P}(w_{i+1} \ldots w_m) z^{m-i}$ denotes the autocorrelation polynomial of w.*

The generating function for the average external profile at level k in a suffix tree is

$$
B(z) = \sum_{\substack{w \in \mathcal{A}^{k-1} \\ \alpha \in \mathcal{A}}} \mathbf{P}(w\alpha) z
$$
$$
\times \left(\frac{1}{D_{w\alpha}(z)^2} - \frac{S_{w\beta}(z) + (1 - S_{w\alpha}(z)) \frac{\mathbf{P}(\beta)}{\mathbf{P}(\alpha)}}{D_w(z)^2} \right).
$$

Proof. The probability that $w\alpha$ appears exactly once in $\mathcal{P}_n$ while $w\beta$ appears at least once in $\mathcal{P}_n$ can be written in an equivalent way that is easier to interpret in terms of combinatorics on words. Notice that $w\alpha$ makes a contribution to the external profile at level k if and only if X begins with a word from $\mathcal{R}_{w\alpha} \mathcal{U}_{w\alpha} \setminus \widetilde{\mathcal{R}}_{w\alpha} \widetilde{\mathcal{U}}_{w\alpha}$. To see this, we note that in $\mathcal{R}_{w\alpha} \mathcal{U}_{w\alpha}$, the word $w\alpha$ appears exactly once. On the other hand, in each word from $\widetilde{\mathcal{R}}_{w\alpha} \widetilde{\mathcal{U}}_{w\alpha}$, we note that $w\alpha$ appears once, but $w\beta$ never appears. For this reason, we must remove the subset $\widetilde{\mathcal{R}}_{w\alpha} \widetilde{\mathcal{U}}_{w\alpha}$ from $\mathcal{R}_{w\alpha} \mathcal{U}_{w\alpha}$. Finally, we restrict attention to words of length $n + k - 1$. So the desired probability is

$$[z^{n+k-1}] \left(R_{w\alpha}(z) U_{w\alpha}(z) - \widetilde{R}_{w\alpha}(z) \widetilde{U}_{w\alpha}(z) \right).$$

It follows that

$$
\sum_{n \geq 0} \mathbb{E}(B_{n,k}) z^n = \sum_{\substack{w \in \mathcal{A}^{k-1} \\ \alpha \in \mathcal{A}}} \frac{1}{z^{k-1}} \left(\frac{\mathbf{P}(w\alpha) z^k}{D_{w\alpha}(z)} \frac{1}{D_{w\alpha}(z)} \right.
$$
$$
\left. - \frac{\mathbf{P}(w\alpha) z^k}{D_w(z)} \frac{S_{w\beta}(z) + (1 - S_{w\alpha}(z)) \frac{\mathbf{P}(\beta)}{\mathbf{P}(\alpha)}}{D_w(z)} \right).
$$

The lemma follows easily by simplifying. ∎

5.3 Generating Function for the Internal Profile of a Trie Constructed from Independent Strings.
We use a simple, direct probabilistic argument to describe the internal profile at level k of a trie constructed from n independent strings $X^{(1)}, \ldots, X^{(n)}$.

We again focus our attention on level k of the trie, so we are primarily concerned with the prefixes of length k of each $X^{(i)}$.

OBSERVATION 3. *The internal profile $\widehat{U}_{n,k}$ at level k of a trie constructed from the n independent strings $X^{(1)}, \ldots, X^{(n)}$ is exactly the sum of the number of words $w \in \mathcal{A}^k$ with the property that $w = X_1^{(i)} \ldots X_k^{(i)}$ for at least two values of i with $1 \leq i \leq n$. In other words, $\widehat{U}_{n,k}$ is the sum of the number of words $w \in \mathcal{A}^k$ that appear as prefixes of at least two of the strings $X^{(1)}, \ldots, X^{(n)}$.*

LEMMA 5.3. *The generating function for the average internal profile at level k in a trie constructed from n independent strings is*

$$
\widehat{U}(z) = \sum_{w \in \mathcal{A}^k} \frac{\mathbf{P}(w)^2 z^2}{(1-z)(1-(1-\mathbf{P}(w))z)^2}.
$$

Proof. The probability that $w \in \mathcal{A}^k$ does not appear as the prefix of any of the strings $X^{(1)}, \ldots, X^{(n)}$ is exactly $(1 - \mathbf{P}(w))^n$. Similarly, w appears as the prefix of exactly one of the strings $X^{(1)}, \ldots, X^{(n)}$ with probability $n\mathbf{P}(w)(1 - \mathbf{P}(w))^{n-1}$. So

$$
\mathbb{E}(\widehat{U}_{n,k}) = \sum_{w \in \mathcal{A}^k} \left(1 - (1 - \mathbf{P}(w))^n \right.
$$
$$
\left. - n\mathbf{P}(w)(1 - \mathbf{P}(w))^{n-1} \right).
$$

Summing $\mathbb{E}(\widehat{U}_{n,k}) z^n$ over all $n \geq 0$, the lemma follows immediately. ∎

5.4 Generating Function for the External Profile of a Trie Constructed from Independent Strings.
We again use a straightforward probabilistic argument in order to establish the external profile at level k of a trie constructed from n independent strings $X^{(1)}, \ldots, X^{(n)}$.

We again focus our attention on level k of the trie, so we are primarily concerned with the prefixes of length k of the various $X^{(i)}$'s. As in our previous discussion of external profiles, we consider words of the form $w\alpha$, where $w \in \mathcal{A}^{k-1}$ and $\alpha \in \mathcal{A}$. We let $\beta = 1 - \alpha$.

Throughout the discussion below, we let $\mathcal{P}_n$ denote the set of prefixes of length k of the strings $X^{(1)}, \ldots, X^{(n)}$.

OBSERVATION 4. *The external profile $\widehat{B}_{n,k}$ at level k of a trie is exactly the sum of the number of words $w\alpha$ for which two conditions hold, namely, $w\alpha$ appears in $\mathcal{P}_n$ exactly once, and also $w\beta$ appears in $\mathcal{P}_n$ at least once. In other words, $\widehat{B}_{n,k}$ is the sum of the number of words $w\alpha$ such that $w\alpha$ appears as a prefix of exactly one of the strings $X^{(1)}, \ldots, X^{(n)}$, and also $w\alpha$'s sibling, namely $w\beta$, appears as the prefix of at least one of the strings $X^{(1)}, \ldots, X^{(n)}$.*

LEMMA 5.4. *The generating function for the average external profile at level k in a trie constructed from n independent strings is*

$$
\widehat{B}(z) = \sum_{\substack{w \in \mathcal{A}^{k-1} \\ \alpha \in \mathcal{A}}} \mathbf{P}(w\alpha) z
$$
$$
\times \left(\frac{1}{(1-(1-\mathbf{P}(w\alpha))z)^2} - \frac{1}{(1-(1-\mathbf{P}(w))z)^2} \right).
$$

Proof. The probability that $w\alpha$ appears exactly once in $\mathcal{P}_n$ while $w\beta$ appears at least once in $\mathcal{P}_n$ can be written in an equivalent—but simpler—way. This is exactly the probability that $w\alpha$ appears exactly once in $\mathcal{P}_n$ (with no restrictions on $w\beta$), minus the probability that $w\alpha$ appears exactly once in $\mathcal{P}_n$ while $w\beta$ does not appear in $\mathcal{P}_n$. We observe that the former is exactly $n\mathbf{P}(w\alpha)(1 - \mathbf{P}(w\alpha))^{n-1}$. We also observe that the latter is exactly the probability that $w\alpha$ appears exactly once in $\mathcal{P}_n$ and w does not appear as the prefix of any of the other $X^{(i)}$'s, namely, the probability $n\mathbf{P}(w\alpha)(1 - \mathbf{P}(w))^{n-1}$. So

$$
\mathbb{E}(\widehat{B}_{n,k}) = \sum_{\substack{w \in \mathcal{A}^{k-1} \\ \alpha \in \mathcal{A}}} n\mathbf{P}(w\alpha)\big((1 - \mathbf{P}(w\alpha))^{n-1}
$$
$$
- (1 - \mathbf{P}(w))^{n-1}\big).
$$

Summing $\mathbb{E}(\widehat{B}_{n,k}) z^n$ over all $n \geq 0$, the lemma follows immediately. ∎

6 Singularity Analysis
We recall that $\mathbf{P}(X_j = 0) = p$ and $\mathbf{P}(X_j = 1) = q$ in the suffix tree model; similarly $\mathbf{P}(X_j^{(i)} = 0) = p$ and $\mathbf{P}(X_j^{(i)} = 1) = q$ in the model with independent

strings. Without loss of generality, we assumed that $0 < q \le p < 1$. We recall that $\delta = \sqrt{p}$; also, $\rho > 1$ is defined such that $\rho\delta < 1$.

We recall from (4.1) the definition of the autocorrelation polynomial of a word $w \in \mathcal{A}^m$ as

$$S_w(z) = \sum_{i \in \mathcal{P}(w)} \mathbf{P}(w_{i+1} \dots w_m) z^{m-i},$$

where $\mathcal{P}(w)$ denotes the set of i's satisfying $w_1 \dots w_i = w_{m-i+1} \dots w_m$. The autocorrelation polynomial $S_w(z)$ records the extent to which w overlaps with itself. Of course, every word w has a trivial (complete) overlap with itself, which provides a contribution of "1" to $S_w(z)$. With high probability, we observe that the other overlaps of w with itself are very small, providing contributions to $S_w(z)$ of much smaller order. We formalize this notion with the following well-known lemma, which appears often throughout the literature of combinatorics on words (see, for instance, [15]). We use the Iverson notation $[\![A]\!] = 1$ if A holds, and $[\![A]\!] = 0$ otherwise.

LEMMA 6.1. *Consider* $\theta = (1 - p\rho)^{-1}$, $\delta = \sqrt{p}$, *and* $\rho > 1$ *with* $\rho\delta < 1$. *When randomly selecting a binary word* $w \in \mathcal{A}^k$, *the autocorrelation polynomial* $S_w(z)$ *(at* $z = \rho$*) is approximately 1, with high probability. More specifically,*

$$\sum_{w \in \mathcal{A}^k} [\![\, |S_w(\rho) - 1| \le (\rho\delta)^k \theta \,]\!] \, \mathbf{P}(w) \ge 1 - \theta\delta^k.$$

LEMMA 6.2. *Recall* $\delta = \sqrt{p}$; *also* $\rho > 1$ *is defined such that* $\rho\delta < 1$. *Consider the polynomial* $D_w(z) = (1-z)S_w(z) + \mathbf{P}(w)z^m$ *associated with a word* $w \in \mathcal{A}^m$, *where* $S_w(z) = \sum_{i \in \mathcal{P}(w)} \mathbf{P}(w_{i+1} \dots w_m) z^{m-i}$ *denotes the autocorrelation polynomial of* w. *There exists an integer* K *such that, for* every *word* w *with* $|w| \ge K$, *the polynomial* $D_w(z)$ *has exactly one root in the disk* $|z| \le \rho$.

Throughout the rest of the discussion below, we fix the "K" mentioned in Lemma 6.2 above, and we consistently restrict attention to word lengths $k \ge K$.

For w with $|w| = k \ge K$, since $D_w(z)$ has a unique root in the disk $|z| \le \rho$, we denote this root as A_w, and we write $B_w = D'_w(A_w)$ and $C_w = D''_w(A_w)$. Using

bootstrapping, we have

$$
\begin{aligned}
A_w &= 1 + \frac{1}{S_w(1)}\mathbf{P}(w) + O(\mathbf{P}(w)^2), \\
B_w &= -S_w(1) + \left(k - \frac{2S'_w(1)}{S_w(1)}\right)\mathbf{P}(w) + O(\mathbf{P}(w)^2), \\
C_w &= -2S'_w(1) + \left(k(k-1) - \frac{3S''_w(1)}{S_w(1)}\right)\mathbf{P}(w) \\
&\quad + O(\mathbf{P}(w)^2).
\end{aligned}
$$

(6.2)

Next we compare $\sum_{n \ge 0} \mathbb{E}(U_{n,k})z^n$ to $\sum_{n \ge 0} \mathbb{E}(\widehat{U}_{n,k})z^n$. Afterwards, using a similar methodology (but omitting some of the details) we compare $\sum_{n \ge 0} \mathbb{E}(B_{n,k})z^n$ to $\sum_{n \ge 0} \mathbb{E}(\widehat{B}_{n,k})z^n$.

7 Comparing Internal Profiles

We define

$$Q(z) = U(z) - \widehat{U}(z) = \sum_{n \ge 0} \left(\mathbb{E}(U_{n,k}) - \mathbb{E}(\widehat{U}_{n,k}) \right) z^n$$

and the contribution to $Q(z)$ from w as

$$Q^{(w)}(z) = \frac{\mathbf{P}(w)z}{1-z}\left(\frac{D_w(z) - (1-z)}{D_w(z)^2} - \frac{\mathbf{P}(w)z}{(1 - (1 - \mathbf{P}(w))z)^2} \right).$$

By Lemmas 5.1 and 5.3, we know that

$$Q(z) = \sum_{w \in \mathcal{A}^k} Q^{(w)}(z).$$

We also define $Q_n = [z^n]Q(z)$ and $Q_n^{(w)} = [z^n]Q^{(w)}(z)$. So Q_n is exactly $\mathbb{E}(U_{n,k}) - \mathbb{E}(\widehat{U}_{n,k})$, and $Q_n^{(w)}$ is the contribution to Q_n from w. Our ultimate goal is to prove that Q_n is asymptotically negligible, i.e., $\mathbb{E}(U_{n,k})$ and $\mathbb{E}(\widehat{U}_{n,k})$ have the same asymptotic growth.

Using Cauchy's Integral Formula, we have

$$Q_n^{(w)} = \frac{1}{2\pi i} \oint \frac{\mathbf{P}(w)z}{1-z}\left(\frac{D_w(z) - (1-z)}{D_w(z)^2} - \frac{\mathbf{P}(w)z}{(1 - (1 - \mathbf{P}(w))z)^2} \right) \frac{dz}{z^{n+1}},$$

where the path of integration is a circle about the origin with counterclockwise orientation. Using a counterclockwise, circular path of radius ρ about the origin, we

also define

$$I_n^{(w)}(\rho) = \frac{1}{2\pi i} \int_{|z|=\rho} \frac{\mathbf{P}(w)z}{1-z} \left(\frac{D_w(z) - (1-z)}{D_w(z)^2} \right.$$

(7.3)
$$\left. - \frac{\mathbf{P}(w)z}{(1-(1-\mathbf{P}(w))z)^2} \right) \frac{dz}{z^{n+1}}.$$

By Cauchy's theorem, it follows that

$$Q_n^{(w)} = I_n^{(w)}(\rho) - \operatorname*{Res}_{z=A_w} \frac{\mathbf{P}(w)z(D_w(z) - (1-z))}{(1-z)D_w(z)^2 z^{n+1}}$$

$$+ \operatorname*{Res}_{z=1/(1-\mathbf{P}(w))} \frac{\mathbf{P}(w)^2 z^2}{(1-z)(1-(1-\mathbf{P}(w))z)^2 z^{n+1}}$$

$$- \operatorname*{Res}_{z=1} \frac{\mathbf{P}(w)z(D_w(z) - (1-z))}{(1-z)D_w(z)^2 z^{n+1}}$$

$$+ \operatorname*{Res}_{z=1} \frac{\mathbf{P}(w)^2 z^2}{(1-z)(1-(1-\mathbf{P}(w))z)^2 z^{n+1}}.$$

We compute the four relevant residues, namely

$$\operatorname*{Res}_{z=A_w} \frac{\mathbf{P}(w)z(D_w(z) - (1-z))}{(1-z)D_w(z)^2 z^{n+1}}$$

$$= \frac{\mathbf{P}(w)}{B_w(1-A_w)A_w^n} + \frac{\mathbf{P}(w)C_w}{B_w^3 A_w^n} + \frac{\mathbf{P}(w)n}{B_w^2 A_w^{n+1}},$$

$$\operatorname*{Res}_{z=1/(1-\mathbf{P}(w))} \frac{\mathbf{P}(w)^2 z^2}{(1-z)(1-(1-\mathbf{P}(w))z)^2 z^{n+1}}$$

(7.4)
$$= (1-\mathbf{P}(w))^{n-1}((n-1)\mathbf{P}(w) + 1),$$

and

$$\operatorname*{Res}_{z=1} \frac{\mathbf{P}(w)z(D_w(z) - (1-z))}{(1-z)D_w(z)^2 z^{n+1}} = -1,$$

$$\operatorname*{Res}_{z=1} \frac{\mathbf{P}(w)^2 z^2}{(1-z)(1-(1-\mathbf{P}(w))z)^2 z^{n+1}} = -1.$$

Now we determine the contribution to $Q_n^{(w)}$ from the first two residues in (7.4). We define

$$f_w(x) = -\frac{\mathbf{P}(w)}{B_w(1-A_w)A_w^x} - \frac{\mathbf{P}(w)C_w}{B_w^3 A_w^x} - \frac{\mathbf{P}(w)x}{B_w^2 A_w^{x+1}}$$

$$+ (1-\mathbf{P}(w))^{x-1}((x-1)\mathbf{P}(w) + 1).$$

We want to prove that $\sum_{w \in \mathcal{A}^k} f_w(x)$ is asymptotically small. We first observe that $\sum_{w \in \mathcal{A}^k} f_w(x)$ is absolutely convergent for all x. Then we define $\bar{f}_w(x) = f_w(x) - f_w(0)e^{-x}$. Next we utilize the Mellin transform of $\bar{f}_w(x)$. (See [9] and [26] for details about the Mellin transform.) Since $\bar{f}_w(x)$ is exponentially decreasing as $x \to +\infty$, and is $O(x)$ when $x \to 0$, then the Mellin transform of $\bar{f}_w(x)$, namely

$$\bar{f}_w^*(s) = \int_0^\infty \bar{f}_w(x)x^{s-1}\,dx,$$

is well-defined for $\Re(s) > 1$. We have

$$\bar{f}_w^*(s) = -\left(\frac{\mathbf{P}(w)}{B_w(1-A_w)} + \frac{\mathbf{P}(w)C_w}{B_w^3} \right)$$

$$\times \int_0^\infty \left(\frac{1}{A_w^x} - 1 \right) x^{s-1}\,dx$$

$$- \frac{\mathbf{P}(w)}{B_w^2 A_w} \int_0^\infty \frac{x}{A_w^x} x^{s-1}\,dx$$

$$+ \frac{\mathbf{P}(w)}{1-\mathbf{P}(w)} \int_0^\infty (1-\mathbf{P}(w))^x x\, x^{s-1}\,dx$$

$$+ \int_0^\infty ((1-\mathbf{P}(w))^x - 1)x^{s-1}\,dx.$$

Using the well-known properties of the Mellin transform (see, for instance, [9] and [26]), it follows that

$$\bar{f}_w^*(s) = -\left(\frac{\mathbf{P}(w)}{B_w(1-A_w)} + \frac{\mathbf{P}(w)C_w}{B_w^3} \right)$$

$$\times \Gamma(s) \left((\log A_w)^{-s} - 1 \right)$$

$$- \frac{\mathbf{P}(w)}{B_w^2 A_w} (\log A_w)^{-s-1}\Gamma(s+1)$$

$$+ \frac{\mathbf{P}(w)}{1-\mathbf{P}(w)} \left(\log \frac{1}{1-\mathbf{P}(w)} \right)^{-s-1} \Gamma(s+1)$$

$$+ \left(\left(\log \frac{1}{1-\mathbf{P}(w)} \right)^{-s} - 1 \right) \Gamma(s).$$

From the bootstrapped equations for A_w, B_w, and C_w given in (6.2), it follows that

$$\bar{f}_w^*(s) = -\left(1 + O(|w|\mathbf{P}(w)^2) \right.$$

$$\left. + \frac{\mathbf{P}(w)2S_w'(1)}{S_w(1)^3}(1 + O(|w|^2\mathbf{P}(w))) \right) \Gamma(s)$$

$$\times \left(\left(\frac{\mathbf{P}(w)}{S_w(1)} \right)^{-s} (1 + O(\mathbf{P}(w))) - 1 \right)$$

$$- \frac{\mathbf{P}(w)}{S_w(1)^2}(1 + O(|w|\mathbf{P}(w))) \left(\frac{\mathbf{P}(w)}{S_w(1)} \right)^{-s-1}$$

$$\times (1 + O(\mathbf{P}(w)))\Gamma(s+1)$$

$$+ \frac{\mathbf{P}(w)}{1-\mathbf{P}(w)} \mathbf{P}(w)^{-s-1}(1 + O(\mathbf{P}(w)))\Gamma(s+1)$$

$$+ \left(\mathbf{P}(w)^{-s}(1 + O(\mathbf{P}(w))) - 1 \right) \Gamma(s),$$

which simplifies to

$$\bar{f}_w^*(s) = \Gamma(s)\mathbf{P}(w)^{-s} \left(1 - \frac{1}{S_w(1)^{-s}} \right)(1 + O(\mathbf{P}(w)))$$

$$+ \Gamma(s+1)\mathbf{P}(w)^{-s} \left(1 - \frac{1}{S_w(1)^{-s+1}} \right)$$

$$\times (1 + O(|w|\mathbf{P}(w))).$$

Now we define $g^*(s) = \sum_{w \in \mathcal{A}^k} \bar{f}_w^*(s)$. We compute

$$g^*(s) = \sum_{w \in \mathcal{A}^k} \mathbf{P}(w)^{-s} \left(\Gamma(s) \left(1 - \frac{1}{S_w(1)^{-s}} \right) \right.$$
$$\left. + \Gamma(s+1) \left(1 - \frac{1}{S_w(1)^{-s+1}} \right) \right) O(1)$$
$$= \sum_{w \in \mathcal{A}^k} \mathbf{P}(w)^{-s-1} \left(\Gamma(s) \frac{\mathbf{P}(w)(S_w(1)^{-s} - 1)}{S_w(1)^{-s}} \right.$$
$$\left. + \Gamma(s+1) \frac{\mathbf{P}(w)(S_w(1)^{-s+1} - 1)}{S_w(1)^{-s+1}} \right) O(1)$$
$$= (\sup\{q^{-\Re(s)-1}, 1\})^k \delta^k (\Gamma(s) + \Gamma(s+1)) O(1),$$

where the last equality follows from Lemma 6.1, which precisely describes the fact that the autocorrelation polynomial is close to 1 with very high probability. We note that when $s = 0$, the pole at $\Gamma(s)$ is canceled.

We note that there exists $c > 0$ such that $q^{-c}\delta < 1$. So $g^*(s)$ is analytic in $\Re(s) \in (-1, c)$. We choose $\epsilon > 0$ with the property that $0 < \epsilon < c$. Then we have

$$Q_n - I_n(\rho) = \frac{1}{2\pi i} \int_{\epsilon - i\infty}^{\epsilon + i\infty} g^*(s) n^{-s} ds + \sum_{w \in \mathcal{A}^k} f_w(0) e^{-x}.$$

The first term is $O(n^{-\epsilon}) O((q^{-c}\delta)^k)$ since $g^*(s)$ is analytic in the strip $\Re(s) \in (-1, c)$. The second term is $O(e^{-x})$. Finally, Lemma 7.1 (given below) concerning $I_n(\rho)$ allows us to complete the proof of Theorem 3.1. In the statement of Theorem 3.1, we use $\mu = q^{-c}\delta < 1$.

LEMMA 7.1. *Consider* $\delta = \sqrt{p}$, *and* $\rho > 1$ *with* $\rho\delta < 1$. *Recall from (7.3) that*

$$I_n^{(w)}(\rho) = \frac{1}{2\pi i} \int_{|z|=\rho} \frac{\mathbf{P}(w)z}{1-z} \left(\frac{D_w(z) - (1-z)}{D_w(z)^2} \right.$$
$$\left. - \frac{\mathbf{P}(w)z}{(1 - (1 - \mathbf{P}(w))z)^2} \right) \frac{dz}{z^{n+1}},$$

where $D_w(z) = (1-z)S_w(z) + \mathbf{P}(w)z^k$ *for* $w \in \mathcal{A}^k$. *The sum of* $I_n^{(w)}(\rho)$ *over all words* $w \in \mathcal{A}^k$ *is asymptotically negligible, namely*

$$\sum_{w \in \mathcal{A}^k} I_n^{(w)}(\rho) = O(\rho^{-n}) O((\rho\delta)^k).$$

Proof. There exist constants C_1, C_2, and K_2 such that, for all $k \geq K_2$ and all $|z| = \rho$, we have $\frac{1}{|D_w(z)|^2} \leq C_1$ and $\frac{1}{|1-(1-\mathbf{P}(w))z|^2} \leq C_2$ for all w with $|w| = k$. The proof of this useful fact is straightforward and appears, for instance, in [27]. Thus

$$|I_n^{(w)}(\rho)| \leq \frac{2\pi\rho}{2\pi} \left(\frac{1}{C_1} \sup_{|z|=\rho} \left| \frac{\mathbf{P}(w)z}{1-z} (D_w(z) - (1-z)) \right| \right.$$
$$\left. - \frac{1}{C_2} \sup_{|z|=\rho} \left| \frac{\mathbf{P}(w)^2 z^2}{1-z} \right| \right) \frac{1}{\rho^{n+1}}.$$

We note that $|D_w(z) - (1-z)| \leq |1-z||S_w(z) - 1| + |z|^k \mathbf{P}(w) \leq (1+\rho)(S_w(\rho) - 1) + (p\rho)^k$. Finally, using Lemma 6.1, which formalizes the notion that the autocorrelation polynomial is close to 1 with high probability, the result follows. ∎

8 Comparing External Profiles

The comparison of the generating functions for the external profile in a suffix tree and in a trie constructed from independent strings is performed in the same way as in the previous section. Due to space constraints, we omit some of the details here, but we do offer the following observations for the interested reader. We observe that

$$D_{w\alpha}(A_w) = \left(-\frac{S_{w\alpha}(1)}{S_w(1)} + \mathbf{P}(\alpha) \right) \mathbf{P}(w)$$
$$+ O(|w|\mathbf{P}(w)^2),$$
$$D'_{w\alpha}(A_w) = -S_{w\alpha}(1) + \left(k - \frac{2S'_{w\alpha}(1)}{S_w(1)} \right) \mathbf{P}(w)$$
$$+ O(\mathbf{P}(w)^2).$$

The relevant residues are

$$\operatorname*{Res}_{z=A_w} \frac{\mathbf{P}(w\alpha)z}{z^{n+1}} \left(\frac{1}{D_{w\alpha}(z)^2} - \frac{S_{w\beta}(z) + (1 - S_{w\alpha}(z))\frac{\mathbf{P}(\beta)}{\mathbf{P}(\alpha)}}{D_w(z)^2} \right)$$
$$= \mathbf{P}(w) \left(\frac{C_w}{B_w^3 A_w^n} + \frac{n}{B_w^2 A_w^{n+1}} + \frac{\mathbf{P}(\alpha)}{B_w(A_w - 1)A_w^n} \right.$$
$$+ \frac{1}{B_w^3(A_w - 1)^2 A_w^{n+1}} \left(D_{w\alpha}(A_w)A_w B_w(n+1) \right.$$
$$- nD_{w\alpha}(A_w)B_w + A_w(1-A_w)(D'_{w\alpha}(A_w)B_w$$
$$\left. \left. - D_{w\alpha}(A_w)C_w) \right) \right)$$

and

$$\operatorname*{Res}_{z=A_{w\alpha}} \frac{\mathbf{P}(w\alpha)z}{z^{n+1}} \left(\frac{1}{D_{w\alpha}(z)^2} - \frac{S_{w\beta}(z) + (1 - S_{w\alpha}(z))\frac{\mathbf{P}(\beta)}{\mathbf{P}(\alpha)}}{D_w(z)^2} \right)$$
$$= -\mathbf{P}(w\alpha) \left(\frac{C_{w\alpha}}{B_{w\alpha}^3 A_{w\alpha}^n} + \frac{n}{B_{w\alpha}^2 A_{w\alpha}^{n+1}} \right),$$
$$\operatorname*{Res}_{z=1/(1-\mathbf{P}(w))} \frac{\mathbf{P}(w\alpha)}{z^{n+1}} z \left(\frac{1}{(1 - (1 - \mathbf{P}(w\alpha))z)^2} \right.$$
$$\left. - \frac{1}{(1 - (1 - \mathbf{P}(w))z)^2} \right)$$
$$= (1 - \mathbf{P}(w))^{n-1} n\mathbf{P}(w\alpha),$$
$$\operatorname*{Res}_{z=1/(1-\mathbf{P}(w\alpha))} \frac{\mathbf{P}(w\alpha)}{z^{n+1}} z \left(\frac{1}{(1 - (1 - \mathbf{P}(w\alpha))z)^2} \right.$$
$$\left. - \frac{1}{(1 - (1 - \mathbf{P}(w))z)^2} \right)$$
$$= -(1 - \mathbf{P}(w\alpha))^{n-1} n\mathbf{P}(w\alpha).$$

The remainder of the proof of Theorem 3.2 is along the same lines as in the previous section.

Acknowledgments

I thank Gahyun Park and Wojciech Szpankowski for several useful discussions concerning [22]. I also thank three anonymous referees for their very insightful comments and suggestions.

References

[1] A. Apostolico and Z. Galil, editors. *Combinatorial Algorithms on Words*, pages 85–96. Cambridge, 1985.

[2] B. Chauvin, M. Drmota, and J. Jabbour-Hattab. The profile of binary search trees. *Annals of Applied Probability*, 11:1042–1062, 2001.

[3] L. Devroye and H.-K. Hwang. Width and mode of the profile for some random trees of logarithmic height. *Annals of Applied Probability*, 16:886–918, 2006.

[4] M. Drmota. Profile and height of random binary search trees. *J. of the Iranian Statistical Society*, 3:117–138, 2004.

[5] M. Drmota and H.-K. Hwang. Bimodality and phase transitions in the profile variance of random binary search trees. *SIAM J. on Discrete Math*, 19:19–45, 2005.

[6] M. Drmota and H.-K. Hwang. Profile of random trees: correlation and width of random recursive trees and binary search trees. *Advanced in Applied Probability*, 37:321–341, 2005.

[7] J. Fayolle. An average-case analysis of basic parameters of the suffix tree. In M. Drmota, P. Flajolet, D. Gardy, and B. Gittenberger, editors, *Mathematics and Computer Science*, pages 217–227, Vienna, Austria, 2004. Birkhäuser.

[8] J. Fayolle and M. D. Ward. Analysis of the average depth in a suffix tree under a Markov model. In Conrado Martínez, editor, *2005 International Conference on Analysis of Algorithms*, volume AD of *DMTCS Proceedings*, pages 95–104, Barcelona, 2005. Discrete Mathematics and Theoretical Computer Science.

[9] P. Flajolet, X. Gourdon, and P. Dumas. Mellin transforms and asymptotics: Harmonic sums. *Theoretical Computer Science*, 144:3–58, 1995.

[10] M. Fuchs, H.-K. Hwang, and R. Neininger. Profiles of random trees: Limit theorems for random recursive trees and binary search trees. *Algorithmica*, 2005. Accepted for publication.

[11] L. Guibas and A. M. Odlyzko. Periods in strings. *J. Combinatorial Theory*, 30:19–43, 1981.

[12] D. Gusfield. *Algorithms on Strings, Trees, and Sequences*. Cambridge, 1997.

[13] H.-K. Hwang. Profiles of random trees: plane-oriented recursive trees. In Conrado Martínez, editor, *2005 International Conference on Analysis of Algorithms*, volume AD of *DMTCS Proceedings*, pages 193–200, Barcelona, 2005. Discrete Mathematics and Theoretical Computer Science.

[14] P. Jacquet and W. Szpankowski. Autocorrelation on words and its applications: Analysis of suffix trees by string-ruler approach. *Journal of Combinatorial Theory*, A66:237–269, 1994.

[15] P. Jacquet and W. Szpankowski. *Applied Combinatorics on Words*, chapter 7, Analytic Approach to Pattern Matching. Cambridge, 2005. See [19].

[16] S. Lonardi, W. Szpankowski, and M. D. Ward. Error resilient LZ'77 data compression: algorithms, analysis, and experiments. Submitted for publication, 2006.

[17] M. Lothaire. *Combinatorics on Words*. Cambridge, 2nd edition, 1997.

[18] M. Lothaire. *Algebraic Combinatorics on Words*. Cambridge, 2002.

[19] M. Lothaire. *Applied Combinatorics on Words*. Cambridge, 2005.

[20] P. Nicodème. Average profiles, from tries to suffixtrees. In Conrado Martínez, editor, *2005 International Conference on Analysis of Algorithms*, volume AD of *DMTCS Proceedings*, pages 257–266, Barcelona, 2005. Discrete Mathematics and Theoretical Computer Science.

[21] G. Park. *Profile of Tries*. PhD thesis, Purdue University, May 2006.

[22] G. Park, H.-K. Hwang, P. Nicodème, and W. Szpankowski. Profiles of tries. Submitted for publication, 2006.

[23] G. Park and W. Szpankowski. Towards a complete characterization of tries. In *SIAM-ACM Symposium on Discrete Algorithms*, Vancouver, 2005.

[24] M. Régnier and A. Denise. Rare events and conditional events on random strings. *Discrete Mathematics and Theoretical Computer Science*, 6:191–214, 2004.

[25] M. Régnier and W. Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, 22:631–649, 1998.

[26] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. Wiley, New York, 2001.

[27] M. D. Ward. *Analysis of an Error Resilient Lempel-Ziv Algorithms Via Suffix Trees*. PhD thesis, Purdue University, May 2005.