Analysis of the multiplicity matching parameter in suffix trees

Mark Daniel Ward[1] and Wojciech Szpankowski[2†]

[1] *Department of Mathematics, Purdue University, West Lafayette, IN, USA.* mward@math.purdue.edu
[2] *Department of Computer Science, Purdue University, West Lafayette, IN, USA.* spa@cs.purdue.edu

In a suffix tree, the multiplicity matching parameter (MMP) M_n is the number of leaves in the subtree rooted at the branching point of the $(n+1)$st insertion. Equivalently, the MMP is the number of pointers into the database in the Lempel-Ziv '77 data compression algorithm. We prove that the MMP asymptotically follows the logarithmic series distribution plus some fluctuations. In the proof we compare the distribution of the MMP in suffix trees to its distribution in tries built over independent strings. Our results are derived by both probabilistic and analytic techniques of the analysis of algorithms. In particular, we utilize combinatorics on words, bivariate generating functions, pattern matching, recurrence relations, analytical poissonization and depoissonization, the Mellin transform, and complex analysis.

Keywords: suffix trees, combinatorics on words, pattern matching, autocorrelation polynomial, complex asymptotics, data compression

1 Introduction

When transmitting data, the goal of *source coding (data compression)* is to represent the source with a minimum of symbols. On the other hand, the goal of *channel coding (error correction)* is to represent the source with a minimum of error probability in decoding. These goals are obviously in conflict. Traditionally, additional symbols are transmitted when performing error correction.

In Lonardi and Szpankowski (2003), an algorithm for joint data compression and error correction is presented; the compression performance is not degraded because the algorithm requires no extra symbols for error correction. In this scheme, a Reed-Solomon error-correcting code is embedded into the Lempel-Ziv '77 data compression algorithm (see Ziv and Lempel (1977)). Lonardi and Szpankowski utilize the fact that the LZ'77 adaptive data compression algorithm is unable to remove all redundancy from the source. Our goal here is to precisely determine the number of redundant bits that are available to be utilized in the aforementioned scheme.

We recall the basic operation of the LZ'77 data compression algorithm. When n bits of the source have already been compressed, the LZ'77 encoder finds the longest prefix of the uncompressed data that also appears in the database (namely, the compressed portion of the data). The encoder performs the compression by storing a pointer into the database (and also the length of this prefix, as well as the next character of the source). Often, this longest prefix appears more than once in the database. Each of the database entries are *equally eligible* for use by the encoder; thus, *any of the analogous pointers* into the database is suitable. In practice, the choice of pointer among these candidates has no significance. On the other hand, by judiciously selecting the pointer, some error correction can be performed. For instance, if two pointers are available, the encoder could easily perform a parity check by choosing the first pointer for "0" and the second pointer for "1". Lonardi and Szpankowski's scheme for performing error correction is very elaborate. We refer the reader to their paper for more details.

We let M_n denote the *number of pointers into the database* when n bits have already been compressed (as described above). Throughout this paper, we are primarily interested in precisely determining the asymptotics of M_n. A thorough analysis of M_n yields a characterization of the degree to which error

[†]This research was supported by NSF Grant CCR-0208709, NIH grant R01 GM068959-01, and AFOSR Grant FA8655-04-1-3074.

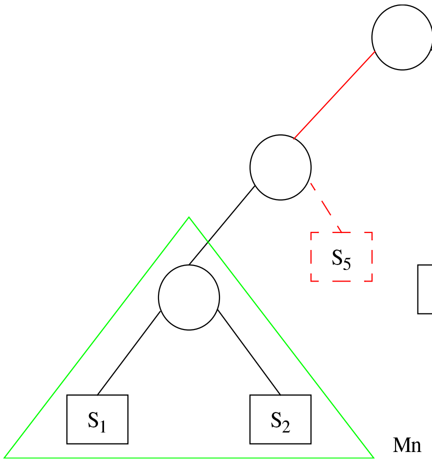

Fig. 1: Finding M_n^I in a trie.

correction can be performed in the scheme discussed above. We note that $\lfloor \log_2 M_n \rfloor$ bits are available to be used for correcting errors.

Tries, especially suffix trees, provide a natural way to study M_n. We work here with strings of characters drawn independently from the binary alphabet $\mathcal{A} := \{0, 1\}$. We let p denote the probability of "0" and $q = 1 - p$ denotes the probability of "1"; without loss of generality, we assume that $q \leq p$ throughout the discussion.

We first recall the definition of a binary trie built over a set $\mathcal{Y}$ of n strings. The construction is recursive. If $|\mathcal{Y}| = 0$, then the trie is empty. If $|\mathcal{Y}| = 1$, then $\mathrm{trie}(\mathcal{Y})$ is a single node. Finally, if $|\mathcal{Y}| > 1$, then $\mathcal{Y}$ is partitioned into two subsets, $\mathcal{Y}_0$ and $\mathcal{Y}_1$, such that a string is in $\mathcal{Y}_0$ if its first symbol is 0, and a string is in $\mathcal{Y}_1$ if its first symbol is 1. Then $\mathrm{trie}(\mathcal{Y}_0)$ and $\mathrm{trie}(\mathcal{Y}_1)$ are each constructed in the same way, except that the splitting of sets at the kth step is based on the kth symbol of the string. This completes the definition of a binary trie.

Now we briefly recall the construction of a binary suffix tree built over a string $X = X_1X_2X_3\ldots$. The word $X^{(i)} = X_iX_{i+1}X_{i+2}\ldots$ is the ith suffix of X, which begins at the ith position of X. Then a binary suffix tree is precisely a binary trie built over the first n suffixes of X, namely $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$.

In a suffix tree, M_n is exactly the number of leaves in the subtree rooted at the branching point of the $(n+1)$st insertion (cf. Figure 1). The strings in a suffix tree are highly dependent on each other, which apparently makes a precise analysis of M_n quite difficult; therefore, we also consider the analogous (but simpler) situation in a trie built over *independent* strings; namely, we study M_n^I, which is the number of leaves in the subtree rooted at the branching point of the $(n+1)$st insertion in a trie built over $n+1$ *independent* binary strings. For instance, in Figure 1, we have $n = 4$ and $M_n^I = 2$ because there are two leaves (namely, S_1 and S_2) in the subtree rooted at the branching point of the 5th insertion.

We are primarily concerned with comparing the distribution of M_n (a parameter of suffix trees) to the distribution of M_n^I (a parameter of tries built over independent strings). Our approach to the proof begins with the observation that a variety of parameters have the same asymptotic behavior regardless of whether they correspond to suffix trees or to tries built over independent strings. This was observed in Szpankowski (1993) and then made precise in Jacquet and Szpankowski (1994), where the typical depth in a suffix tree is proven to be asymptotically the same as the typical depth in a trie built over independent strings when the underlying source is i.i.d. An extension of such results to an underlying Markovian model is presented in Fayolle and Ward (2005).

The limiting distribution of several trie parameters is given in Jacquet and Régnier (1986) and Jacquet and Régnier (1987). More results about trie parameters are found in Kirschenhofer and Prodinger (1988). The variance of the external path in a symmetric trie is given in Kirschenhofer et al. (1989). The depth of a digital trie with an underlying Markovian dependency is analyzed in Jacquet and Szpankowski (1991). Many results about a variety of tree structures are collected in Szpankowski (2001). Average-case studies of several parameters of suffix trees are found in Fayolle (2004) and Szpankowski (1993).

We briefly summarize the methodology of our proof. Our goal is to compare the distribution of M_n (the multiplicity matching parameter of a suffix tree) to the distribution of M_n^I (the MMP of a trie built over independent strings). Our proof that these two parameters have the same asymptotic distribution consists of several steps. We first derive bivariate generating functions for M_n and M_n^I, denoted as $M(z,u)$ and $M^I(z,u)$, respectively. We noted above that a suffix tree is built over the suffixes $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ of a string X. These suffixes are highly dependent on each other. Therefore, in deriving the bivariate generating function $M(z,u)$, an interesting obstacle arises: We need to determine the degree to which a suffix of X can overlap with itself. Fortunately, the autocorrelation polynomial $S_w(z)$ of a word w measures the amount of overlap of a word w with itself. The autocorrelation polynomial was introduced in Guibas and Odlyzko (1981) and was utilized extensively in Régnier and Szpankowski (1998) and Lothaire (2005). The autocorrelation polynomial is defined as

$$S_w(z) = \sum_{k \in \mathcal{P}(w)} \mathbf{P}(w_{k+1}^m) z^{m-k} \tag{1}$$

where $m = |w|$ and where $\mathcal{P}(w)$ denotes the set of positions k of w satisfying $w_1 \ldots w_k = w_{m-k+1} \ldots w_m$, that is, w's prefix of length k is equal to w's suffix of length k. Using the autocorrelation polynomial, we can overcome the difficulties inherent in the fact that suffixes of a word X overlap with each other. By utilizing $S_w(z)$, we are able to obtain a succinct way of describing the bivariate generating function $M(z,u)$. Fortunately, the autocorrelation polynomial is well-understood. Note that the autocorrelation polynomial $S_w(z)$ has a $\mathbf{P}(w_{k+1}^m)z^{m-k}$ term if and only if w has an overlap with itself of length k. All words w overlap with themselves trivially, so all autocorrelation polynomials have a constant term (i.e., $z^{m-m} = z^0 = 1$ term). On the other hand, with high probability, w has very few large nontrivial overlaps with itself. Therefore, with high probability, all nontrivial overlaps of w with itself are small; such overlaps correspond to high-degree terms of $S_w(z)$.

In order to compare $M(z,u)$ and $M^I(z,u)$, we utilize complex analysis. Specifically, we take advantage of Cauchy's theorem, which allows us to analyze the poles of the generating functions $M(z,u)$ and $M^I(z,u)$ in order to obtain precise information about the distributions of M_n and M_n^I. During this residue analysis, it is necessary that the generating function for M_n is analytically continued from the unit disk to a larger disk.

Our ultimate conclusion is that the distribution of the multiplicity matching parameter M_n is asymptotically the same in suffix trees and independent tries, i.e., M_n and M_n^I have asymptotically the same distribution.

The asymptotics for the distribution and factorial moments of M_n^I were given in Ward and Szpankowski (2004). Specifically, M_n^I asymptotically follows the logarithmic series distribution (plus some fluctuations when $\ln p / \ln q$ is rational). Since we prove here that M_n and M_n^I have asymptotically the same distribution, then as a consequence, we see that M_n also asymptotically follows the logarithmic series distribution. One striking property of this distribution is the high concentration around the mean. We see that $E[M_n]$ is asymptotically $\frac{1}{h}$ (where h denotes the entropy of the source) and also M_n is highly concentrated around this average value; this property of M_n is very desirable for the error correction scheme described in Lonardi and Szpankowski (2003).

This paper is a concise version of the first author's Ph.D. thesis; see Ward (2005).

2 Main Results

We consider the string $X = X_1X_2X_3\ldots$, where the X_i's are i.i.d. random variables on $\mathcal{A} := \{0,1\}$ with $\mathbf{P}(X_i = 0) = p$ and $\mathbf{P}(X_i = 1) = q$. (Without loss of generality, we assume throughout the discussion that $q \le p$.) Let $X^{(i)}$ denote the ith suffix of X. In other words, $X^{(i)} = X_iX_{i+1}X_{i+2}\ldots$. Consider the *longest* prefix w of $X^{(n+1)}$ such that $X^{(i)}$ also has w as a prefix, for some i with $1 \le i \le n$. Then M_n is defined as the number of $X^{(i)}$'s (with $1 \le i \le n$) that have w as a prefix. So

$$M_n = \#\{1 \le i \le n \mid X^{(i)} = X_iX_{i+1}X_{i+2}\ldots \text{ has } w \text{ as a prefix}\}\,. \tag{2}$$

An alternate definition of M_n is available via *suffix trees*. First, consider a suffix tree built from the first $n+1$ suffixes of X. Next, consider the *insertion point* of the $(n+1)$st suffix. Then M_n is exactly the *number of leaves* in the subtree rooted at the branching point of the $(n+1)$st insertion. For instance, suppose that the $(n+1)$st suffix starts with $w\beta$ for some $w \in \mathcal{A}^*$ and $\beta \in \mathcal{A}$. Then, examining the first n suffixes, if there are exactly k suffixes that begin with $w\alpha$ (where $\alpha = 1-\beta$, i.e., $\{\alpha,\beta\} = \{0,1\}$), and the other $n-k$ suffixes do not begin with w, we conclude that $M_n = k$.

Unfortunately, the strings in a suffix tree are highly dependent on each other; thus, a precise analysis of M_n is quite difficult. On the other hand, the asymptotic behavior of M_n^I, an analogous parameter of tries built over independent strings, is well-understood. Specifically, M_n^I asymptotically follows the logarithmic series distribution (plus some fluctuations when $\ln p/\ln q$ is rational). In Ward and Szpankowski (2004), a precise analysis of M_n^I is given via the analysis of independent tries, using recurrence relations, analytical poissonization and depoissonization, the Mellin transform, and complex analysis.

To define M_n^I, we consider the situation described above, but we build a trie from $n+1$ independent strings from $\mathcal{A}^*$. So we consider independent $X(i)$'s; specifically, we define $X(i) = X_1(i)X_2(i)X_3(i)\ldots$, where $\{X_j(i) \mid i, j \in \mathbb{N}\}$ is a collection of i.i.d. random variables. We let w denote the *longest* prefix of $X(n+1)$ such that $X(i)$ also has w as a prefix, for some i with $1 \le i \le n$. Then M_n^I is defined as the number of $X(i)$'s (with $1 \le i \le n$) that have w as a prefix. So

$$M_n^I = \#\{1 \le i \le n \mid X(i) = X_1(i)X_2(i)X_3(i)\ldots \text{ has } w \text{ as a prefix}\}\,. \tag{3}$$

To define M_n^I via *tries*, first consider a trie built from the $n+1$ independent strings from $\mathcal{A}^*$. Next, consider the *insertion point* of the $(n+1)$st string. Then M_n is exactly the *number of leaves* in the subtree rooted at the branching point of the $(n+1)$st insertion. As above, suppose that the $(n+1)$st string starts with $w\beta$. Then, examining the first n strings, if there are exactly k strings that begin with $w\alpha$ (again $\alpha = 1-\beta$), and the other $n-k$ strings do not begin with w, we conclude that $M_n^I = k$.

Since we know from Ward and Szpankowski (2004) that M_n^I follows the logarithmic series distribution plus some fluctuations, then it suffices to prove that M_n has a similar asymptotic distribution. To accomplish this goal, we compare the distribution of M_n in suffix trees to the distribution of M_n^I in independent tries.

Briefly, our proof technique is the following: We let $M(z,u) = \sum_{1\le k,n\le\infty} \mathbf{P}(M_n = k)u^k z^n$ and $M^I(z,u) = \sum_{1\le k,n\le\infty} \mathbf{P}(M_n^I = k)u^k z^n$ denote the bivariate generating functions for M_n and M_n^I, respectively. To study these generating functions, we consider the w's defined above. Specifically, for $M(z,u)$, we recall from (2) that if w denotes the longest prefix of $X^{(n+1)} = X_{n+1}X_{n+2}X_{n+3}\ldots$ that appears as a prefix of any $X^{(i)} = X_iX_{i+1}X_{i+2}\ldots$, then M_n enumerates the number of such occurrences of w. This approach to $M(z,u)$ allows us to sum over all $w \in \mathcal{A}^*$ instead of summing over $k, n \in \mathbb{N}$. Similarly, for $M^I(z,u)$, we utilize (3) to see that if w denotes the longest prefix of $X(n+1) = X_1(n+1)X_2(n+1)X_3(n+1)\ldots$ that appears as a prefix of any $X_1(i)X_2(i)X_3(i)\ldots$, then M_n^I is precisely the number of such occurrences of w. Therefore, to determine $M^I(z,u)$, we can sum over all $w \in \mathcal{A}^*$ instead of summing over the integers k and n.

We note that the $X^{(i)}$'s are highly dependent on each other. In fact, if $i \ge j$, then $X^{(i)} = X_iX_{i+1}X_{i+2}\ldots$ is a substring of $X^{(j)} = X_jX_{j+1}X_{j+2}\ldots$. This apparently makes the derivation of the bivariate generating function $M(z,u)$ quite difficult. We overcome this hurdle by succinctly describing the degree to which a suffix of X can overlap with itself. We accomplish this by utilizing the autocorrelation polynomial $S_w(z)$ of a word w, which measures the amount of overlap of a word w with itself. As mentioned above, the autocorrelation polynomial is defined as

$$S_w(z) = \sum_{k\in\mathcal{P}(w)} \mathbf{P}(w_{k+1}^m)z^{m-k} \tag{4}$$

where $\mathcal{P}(w)$ denotes the set of positions k of w satisfying $w_1\ldots w_k = w_{m-k+1}\ldots w_m$, that is, w's prefix of length k is equal to w's suffix of length k. Via the autocorrelation polynomial, we are able to surmount the difficulties inherent in the overlapping suffixes. Thus, using $S_w(z)$, we obtain a succinct description of the bivariate generating function $M(z,u)$. The autocorrelation polynomial is well-understood; we utilize several results about $S_w(z)$ from Régnier and Szpankowski (1998) and Lothaire (2005). In particular, when comparing $M(z,u)$ and $M^I(z,u)$, it is extremely useful to note that the autocorrelation polynomial $S_w(z)$ is close to 1 with high probability (for $|w|$ large).

In order to obtain information about the difference of the two BGFs as $Q(z,u) = M(z,u) - M^I(z,u)$, we utilize residue analysis. We make a comparison of the poles of $M(z,u)$ and $M^I(z,u)$ using Cauchy's theorem (integrating with respect to z). As a result, we prove that $Q_n(u) := [z^n]Q(z,u) = O(n^{-\epsilon})$ uniformly for $|u| \le p^{-1/2}$ as $n \to \infty$. Then we use another application of Cauchy's theorem (integrating with respect to u). Specifically, we extract the coefficient $\mathbf{P}(M_n = k) - \mathbf{P}(M_n^I = k) = [u^k z^n]Q(z,u)$ in order to prove our main result.

Theorem 2.1 *There exist $\epsilon > 0$ and $b > 1$ such that*

$$\mathbf{P}(M_n = k) - \mathbf{P}(M_n^I = k) = O(n^{-\epsilon}b^{-k}) \tag{5}$$

for large n.

Therefore, the distributions of M_n and M_n^I are asymptotically the same. We conclude that M_n also asymptotically follows the logarithmic series distribution (plus some fluctuations when $\ln p/\ln q$ is rational).

Theorem 2.2 *There exist $\epsilon > 0$ and $\epsilon_j > 0$ (for each $j \in \mathbb{N}$) depending on p such that the jth factorial moment of M_n is*

$$\mathbf{E}[(M_n)^{\underline{j}}] = \Gamma(j)\frac{q(p/q)^j + p(q/p)^j}{h} + \gamma_j(\log_{1/p} n) + O(n^{-\epsilon_j}) \tag{6}$$

where γ_j is a periodic function with mean 0 and small modulus if $\ln p/\ln q$ is rational, and otherwise $\gamma_j(x) \to 0$ as $x \to \infty$. Also $h = -p\log p - q\log q$ denotes the entropy of the source. The probability generating function of M_n is

$$\mathbf{E}[u^{M_n}] = -\frac{q\ln(1-pu) + p\ln(1-qu)}{h} + \gamma(\log_{1/p} n, u) + O(n^{-\epsilon}), \tag{7}$$

for $|u| \le p^{-1/2}$ where $\gamma(\cdot, u)$ is a periodic function with mean 0 and small modulus if $\ln p/\ln q$ is rational, and otherwise $\gamma_j(u,x) \to 0$ (uniformly for $|u| \le p^{-1/2}$) as $x \to \infty$. More precisely,

$$\mathbf{E}[u^{M_n}] = \sum_{j=1}^{\infty}\left[\frac{p^j q + q^j p}{jh} + \sum_{k\in\mathbb{Z}\setminus\{0\}} -\frac{e^{2kr\pi i \log_{1/p} n}\Gamma(z_k)(p^j q + q^j p)(z_k)^{\overline{j}}}{j!(p^{-z_k+1}\ln p + q^{-z_k+1}\ln q)}\right]u^j + O(n^{-\epsilon}) \tag{8}$$

when $\ln p/\ln q = r/t$ for some $r, t \in \mathbb{Z}$, we have $z_k = 2kr\pi i/\ln p$. Therefore, as $n \to \infty$, we conclude that M_n follows the logarithmic series distribution plus some fluctuations if $\ln p/\ln q = r/t$ is rational, i.e.,

$$\mathbf{P}(M_n = j) = \frac{p^j q + q^j p}{jh} + \sum_{k\neq 0} -\frac{e^{2kr\pi i \log_{1/p} n}\Gamma(z_k)(p^j q + q^j p)(z_k)^{\overline{j}}}{j!(p^{-z_k+1}\ln p + q^{-z_k+1}\ln q)} + O(n^{-\epsilon}). \tag{9}$$

If $\ln p/\ln q$ is irrational, then M_n asymptotically follows the logarithmic series distribution, without fluctuations.

Note that the average value of M_n is asymptotically $\frac{1}{h}$, and also M_n is highly concentrated around the mean; this property of M_n is very desirable for the error correction scheme described in Lonardi and Szpankowski (2003).

3 Proofs

We first derive the bivariate generating functions for M_n and M_n^I, denoted as $M(z,u)$ and $M^I(z,u)$, respectively. Then we prove a few useful lemmas concerning the autocorrelation polynomial. Next, we prove that $M(z,u)$ can be analytically continued from the unit disk to a larger disk. Afterwards, we determine the poles of $M(z,u)$ and $M^I(z,u)$. We write $Q(z,u) = M(z,u) - M^I(z,u)$; we use Cauchy's theorem to that $Q_n(u) := [z^n]Q(z,u) \to 0$ uniformly for $u \le p^{-1/2}$ as $n \to \infty$. Then we apply Cauchy's theorem again to prove that $\mathbf{P}(M_n = k) - \mathbf{P}(M_n^I = k) = [u^k z^n]Q(z,u) = O(n^{-\epsilon}b^{-k})$ for some $\epsilon > 0$ and $b > 1$.

We conclude that the distribution of the multiplicity matching parameter M_n is asymptotically the same in suffix trees as in tries built over independent strings, i.e., M_n and M_n^I have asymptotically the same distribution. Therefore, M_n also follows the logarithmic series distribution plus some fluctuations.

3.1 BGF for the Multiplicity Matching Parameter of Independent Tries

First we obtain the bivariate generating function for M_n^I, which is the multiplicity matching parameter for a trie built over the *independent* strings $X(1), \ldots, X(n+1)$, where $X(i) = X_1(i)X_2(i)X_3(i)\ldots$ and $\{X_j(i) \mid i, j \in \mathbb{N}\}$ is a collection of i.i.d. random variables with $\mathbf{P}(X_j(i) = 0) = p$ and $\mathbf{P}(X_j(i) = 1) = q = 1-p$. We let w denote the *longest prefix* of both $X(n+1)$ and at least one other string $X(i)$ for some $1 \le i \le n$. We write β to denote the $(|w|+1)$st character of $X(n+1)$. When $M_n^I = k$, we conclude that exactly k strings $X(i)$ have $w\alpha$ as a prefix, and the other $n-k$ strings $X(i)$ do not have w as a prefix at all. Thus the generating function for M_n^I is exactly

$$M^I(z,u) := \sum_{n=1}^{\infty}\sum_{k=1}^{\infty}\mathbf{P}(M_n^I = k)u^k z^n = \sum_{n=1}^{\infty}\sum_{k=1}^{\infty}\sum_{\substack{w\in\mathcal{A}^*\\ \alpha\in\mathcal{A}}}\mathbf{P}(w\beta)\binom{n}{k}(\mathbf{P}(w\alpha))^k(1-\mathbf{P}(w))^{n-k}u^k z^n. \tag{10}$$

After simplifying, it follows immediately that

$$M^I(z,u) = \sum_{\substack{w\in\mathcal{A}^*\\ \alpha\in\mathcal{A}}} \frac{u\mathbf{P}(\beta)\mathbf{P}(w)}{1-z(1-\mathbf{P}(w))}\frac{z\mathbf{P}(w)\mathbf{P}(\alpha)}{1-z(1+u\mathbf{P}(w)\mathbf{P}(\alpha)-\mathbf{P}(w))}\,. \tag{11}$$

Our reasoning about $M^I(z,u)$ can be applied when we derive generating function $M(z,u)$ for M_n in the next section, but the situation will be more complicated, because the occurrences of w can overlap.

3.2 BGF for the Multiplicity Matching Parameter of Suffix Trees

Now we obtain the bivariate generating function for M_n, which is the multiplicity matching parameter for a suffix tree built over the first $n+1$ *suffixes* $X^{(1)},\ldots,X^{(n+1)}$ of a string X (i.e., $X^{(i)} = X_iX_{i+1}X_{i+2}\ldots$). The bivariate generating function for the multiplicity matching parameter is much more difficult to derive in the dependent (suffix tree) case than in the independent (trie) case, because the suffixes of X are dependent on each other. We let w denote the *longest prefix* of both $X^{(n+1)}$ and at least one $X^{(i)}$ for some $1\le i\le n$. We write β to denote the $(|w|+1)$st character of $X^{(n+1)}$; when $M_n=k$, we conclude that exactly k suffixes $X^{(i)}$ have $w\alpha$ as a prefix, and the other $n-k$ strings $X^{(i)}$ do not have w as a prefix at all. Thus, we are interested in finding strings with exactly k occurrences of $w\alpha$, ended on the right by an occurrence of $w\beta$, with no other occurrences of w at all. This set of words is exactly the language $\mathcal{R}_w\alpha(\mathcal{T}_w^{(\alpha)}\alpha)^{k-1}\mathcal{T}_w^{(\alpha)}\beta$, where

$$\begin{aligned}\mathcal{R}_w &= \{v \mid v \text{ contains exactly one occurrence of } w,\ \text{located at the right end}\}\\ \mathcal{T}_w^{(\alpha)} &= \{v \mid w\alpha v \text{ contains exactly two occurrences of } w,\ \text{located at the left and right ends}\}\end{aligned} \tag{12}$$

So, the generating function for M_n is

$$M(z,u) = \sum_{k=1}^{\infty}\sum_{\substack{w\in\mathcal{A}^*\\ \alpha\in\mathcal{A}}}\sum_{s\in\mathcal{R}_w}\mathbf{P}(s\alpha)z^{|s|+1}u\Bigg(\sum_{t\in\mathcal{T}_w^{(\alpha)}}\mathbf{P}(t\alpha)z^{|t|+1}u\Bigg)^{k-1}\sum_{v\in\mathcal{T}_w^{(\alpha)}}\mathbf{P}(v\beta)z^{|v|+1-|w|-1}\,. \tag{13}$$

After simplifying the geometric sum, this yields

$$M(z,u) = \sum_{\substack{w\in\mathcal{A}^*\\ \alpha\in\mathcal{A}}} u\mathbf{P}(\beta)\frac{R_w(z)}{z^{|w|}}\frac{\mathbf{P}(\alpha)zT_w^{(\alpha)}(z)}{1-\mathbf{P}(\alpha)zuT_w^{(\alpha)}(z)}\,. \tag{14}$$

We note that $R_w(z)/z^{|w|} = \mathbf{P}(w)/D_w(z)$ (Régnier and Szpankowski (1998)), where $D_w(z) = (1-z)S_w(z)+z^m\mathbf{P}(w)$ and where $S_w(z)$ denotes the autocorrelation polynomial for w. Recall that $S_w(z)$ measures the degree to which a word w overlaps with itself, and specifically

$$S_w(z) = \sum_{k\in\mathcal{P}(w)}\mathbf{P}(w_{k+1}^m)z^{m-k} \tag{15}$$

where $\mathcal{P}(w)$ denotes the set of positions k of w satisfying $w_1\ldots w_k = w_{m-k+1}\ldots w_m$, that is, w's prefix of length k is equal to w's suffix of length k; also, $m=|w|$. Returning to (14), it follows that

$$M(z,u) = \sum_{\substack{w\in\mathcal{A}^*\\ \alpha\in\mathcal{A}}} \frac{u\mathbf{P}(\beta)\mathbf{P}(w)}{D_w(z)}\frac{\mathbf{P}(\alpha)zT_w^{(\alpha)}(z)}{1-\mathbf{P}(\alpha)zuT_w^{(\alpha)}(z)}\,. \tag{16}$$

In order to derive an explicit form of $M(z,u)$, we still need to find $T_w^{(\alpha)}(z)$. If we define

$$\mathcal{M}_w = \{v \mid wv \text{ contains exactly two occurrences of } w,\ \text{located at the left and right ends}\} \tag{17}$$

then we observe that $\alpha\mathcal{T}_w^{(\alpha)}$ is exactly the subset of words of $\mathcal{M}_w$ that begin with α; We use $\mathcal{H}_w^{(\alpha)}$ to denote this subset (i.e., $\mathcal{H}_w^{(\alpha)} = \mathcal{M}_w\cap(\alpha\mathcal{A}^*)$), and thus $\alpha\mathcal{T}_w^{(\alpha)} = \mathcal{H}_w^{(\alpha)}$. So (16) simplifies to

$$M(z,u) = \sum_{\substack{w\in\mathcal{A}^*\\ \alpha\in\mathcal{A}}} \frac{u\mathbf{P}(\beta)\mathbf{P}(w)}{D_w(z)}\frac{H_w^{(\alpha)}(z)}{1-uH_w^{(\alpha)}(z)}\,. \tag{18}$$

In order to compute $H_w^{(\alpha)}(z)$, we write $\mathcal{M}_w = \mathcal{H}_w^{(\alpha)} + \mathcal{H}_w^{(\beta)}$, where $\mathcal{H}_w^{(\beta)}$ is the subset of words from $\mathcal{M}_w$ that start with β (i.e., $\mathcal{H}_w^{(\beta)} = \mathcal{M}_w \cap (\beta\mathcal{A}^*)$). (Note that every word of $\mathcal{M}_w$ begins with either α or β, because the empty word $\varepsilon \notin \mathcal{M}_w$.) The following useful lemma is the last necessary ingredient to obtain an explicit formula for $M(z,u)$ from (18).

Lemma 3.1 *Let $\mathcal{H}_w^{(\alpha)}$ denote the subset of words from $\mathcal{M}_w$ that start with α. Then*

$$H_w^{(\alpha)}(z) = \frac{D_{w\alpha}(z) - (1-z)}{D_w(z)}\,. \tag{19}$$

Proof We use the concepts and notation from Régnier and Szpankowski (1998) and Lothaire (2005) throughout. In particular, we define

$$\mathcal{U}_w = \{v \mid wv \text{ contains exactly one occurrence of } w \text{ (located at the left end)}\} \tag{20}$$

and we recall from (12) and (17) above that

$$\begin{aligned} \mathcal{R}_w &= \{v \mid v \text{ contains exactly one occurrence of } w, \text{ located at the right end}\} \\ \mathcal{M}_w &= \{v \mid wv \text{ contains exactly two occurrences of } w, \text{ located at the left and right ends}\} \end{aligned} \tag{21}$$

The following notation is similar but slightly adapted for our proof.

$$\mathcal{U}_w^{(\alpha)} = \{v \mid v \text{ starts with } \alpha, \text{ and } wv \text{ has exactly 1 occurrence of } w\alpha \text{ and no occurrences of } w\beta\}\,. \tag{22}$$

We note that the set of words with no occurrences of $w\beta$ is exactly $\mathcal{A}^* \setminus \mathcal{R}_{w\beta}(\mathcal{M}_{w\beta})^*\mathcal{U}_{w\beta}$, which has generating function

$$\frac{1}{1-z} - \frac{R_{w\beta}(z)U_{w\beta}(z)}{1-M_{w\beta}(z)}\,. \tag{23}$$

Now we describe the set of words with no occurrences of $w\beta$ in a different way. The set of words with no occurrences of $w\beta$ and at least one occurrence of $w\alpha$ is exactly $\mathcal{R}_w(\mathcal{H}_w^{(\alpha)})^*\mathcal{U}_w^{(\alpha)}$, which has generating function $R_w(z)U_w^{(\alpha)}(z)/(1-H_w^{(\alpha)}(z))$. The set of words with no occurrences of $w\beta$ and no occurrences of $w\alpha$ is exactly $\mathcal{R}_w + (\mathcal{A}^* \setminus \mathcal{R}_w(\mathcal{M}_w)^*\mathcal{U})$. (Note that the set of such words that end in w is exactly $\mathcal{R}_w$; on the other hand, the set of such words that do not end in w is exactly $\mathcal{A}^* \setminus \mathcal{R}_w(\mathcal{M}_w)^*\mathcal{U}$.) So the set of words with no occurrences of $w\alpha$ and no occurrences of $w\beta$ has generating function $R_w(z) + 1/(1-z) - R_w(z)U_w(z)/(1-M_w(z))$. So the set of words with no occurrences of $w\beta$ has generating function

$$\frac{R_w(z)U_w^{(\alpha)}(z)}{1-H_w^{(\alpha)}(z)} + R_w(z) + \frac{1}{1-z} - \frac{R_w(z)U_w(z)}{1-M_w(z)}\,. \tag{24}$$

Combining (23) and (24), it follows that

$$\frac{1}{1-z} - \frac{R_{w\beta}(z)U_{w\beta}(z)}{1-M_{w\beta}(z)} = \frac{R_w(z)U_w^{(\alpha)}(z)}{1-H_w^{(\alpha)}(z)} + R_w(z) + \frac{1}{1-z} - \frac{R_w(z)U_w(z)}{1-M_w(z)}\,. \tag{25}$$

Now we find the generating function for $\mathcal{U}_w^{(\alpha)}$. For each word $v \in \mathcal{U}_w^{(\alpha)}$, either wv has exactly one or two occurrences of w. The subset of $\mathcal{U}_w^{(\alpha)}$ of the first type is exactly $V_w^{(\alpha)} := U_w \cap (\alpha\mathcal{A}^*)$, i.e., the subset of words from U_w that start with α. The subset of $\mathcal{U}_w^{(\alpha)}$ of the second type is exactly $\mathcal{H}_w^{(\alpha)}$. We observe that

$$\mathcal{V}_w^{(\alpha)} \cdot \mathcal{A} = (\mathcal{H}_w^{(\alpha)} + \mathcal{V}_w^{(\alpha)}) \setminus \{\alpha\} \tag{26}$$

(see Ward (2005)), so $V_w^{(\alpha)}(z) = (H_w^{(\alpha)}(z) - \mathbf{P}(\alpha)z)/(z-1)$. Since $\mathcal{U}_w^{(\alpha)} = \mathcal{V}_w^{(\alpha)} + \mathcal{H}_w^{(\alpha)}$, it follows that

$$U_w^{(\alpha)}(z) = \frac{H_w^{(\alpha)}(z) - \mathbf{P}(\alpha)z}{z-1} + H_w^{(\alpha)}(z) = \frac{zH_w^{(\alpha)}(z) - \mathbf{P}(\alpha)z}{z-1}\,. \tag{27}$$

Recalling equation (25), we see that

$$\frac{1}{1-z} - \frac{R_{w\beta}(z)U_{w\beta}(z)}{1-M_{w\beta}(z)} = \frac{R_w(z)(zH_w^{(\alpha)}(z) - \mathbf{P}(\alpha)z)}{(1-H_w^{(\alpha)}(z))(z-1)} + R_w(z) + \frac{1}{1-z} - \frac{R_w(z)U_w(z)}{1-M_w(z)}\,. \tag{28}$$

Simplifying, and using $U_w(z) = (1-M_w(z))/(1-z)$ and $U_{w\beta}(z) = (1-M_{w\beta}(z))/(1-z)$ (see Régnier and Szpankowski (1998)), it follows that

$$\frac{R_{w\beta}(z)}{R_w(z)} = \frac{z\mathbf{P}(\beta)}{1 - H_w^{(\alpha)}(z)}\,. \tag{29}$$

Solving for $H_w^{(\alpha)}(z)$ and then using $R_w(z) = z^m\mathbf{P}(w)/D_w(z)$ and $R_{w\beta}(z) = z^{m+1}\mathbf{P}(w)\mathbf{P}(\beta)/D_{w\beta}(z)$ (see Régnier and Szpankowski (1998)), it follows that

$$H_w^{(\alpha)}(z) = \frac{D_w(z) - D_{w\beta}(z)}{D_w(z)}\,. \tag{30}$$

Note $D_w(z)-D_{w\beta}(z) = (1-z)S_w(z)+z^m\mathbf{P}(w)-(1-z)S_{w\beta}(z)-z^{m+1}\mathbf{P}(w)\mathbf{P}(\beta) = (1-z)(S_{w\alpha}(z)-1) + z^{m+1}\mathbf{P}(w)\mathbf{P}(\alpha) = D_{w\alpha}(z) - (1-z)$. Thus, (30) completes the proof of the lemma. □

Using the lemma above, we finally observe a form of $M(z,u)$ that we summarize below.

Theorem 3.1 *Let $M(z,u) := \sum_{n=1}^{\infty}\sum_{k=1}^{\infty}\mathbf{P}(M_n = k)u^k z^n$ denote the bivariate generating function for M_n, the multiplicity matching parameter of a suffix tree built over the first $n+1$ suffixes $X^{(1)},\dots,X^{(n+1)}$ of a string X. Then*

$$M(z,u) = \sum_{\substack{w\in\mathcal{A}^*\\ \alpha\in\mathcal{A}}} \frac{u\mathbf{P}(\beta)\mathbf{P}(w)}{D_w(z)}\,\frac{D_{w\alpha}(z)-(1-z)}{D_w(z) - u(D_{w\alpha}(z)-(1-z))} \tag{31}$$

for $|u| < 1$ and $|z| < 1$. Here $D_w(z) = (1-z)S_w(z) + z^m\mathbf{P}(w)$, and $S_w(z)$ denotes the autocorrelation polynomial for w, defined in (1).

3.3 On the Autocorrelation Polynomial

Throughout the rest of our analysis we assume that, without loss of generality, $p \ge q$. Note that $p \le \sqrt{p} < 1$, so there exists $\rho > 1$ such that $\rho\sqrt{p} < 1$ (and thus $\rho p < 1$ too). Finally, define $\delta = \sqrt{p}$. We establish a few lemmas about the autocorrelation polynomial that will be important for our analysis. Recall that the autocorrelation polynomial is $S_w(z) = \sum_{k\in\mathcal{P}(w)}\mathbf{P}(w_{k+1}^m)z^{m-k}$, where $\mathcal{P}(w)$ denotes the set of positions k of w satisfying $w_1\dots w_k = w_{m-k+1}\dots w_m$, that is, w's prefix of length k is equal to w's suffix of length k.

The autocorrelation polynomial $S_w(z)$ has a $\mathbf{P}(w_{k+1}^m)z^{m-k}$ term if and only if w has an overlap with itself of length k. Since each word w overlaps with itself trivially, then every autocorrelation polynomial has a constant term (i.e., $z^{m-m} = z^0 = 1$ term). With high probability, however, w has very few large nontrivial overlaps with itself. Therefore, with high probability, all nontrivial overlaps of w with itself are small; such overlaps correspond to high-degree terms of $S_w(z)$. Therefore, when w is a randomly chosen long word, then $S_w(z)$ is very close to 1 with very high probability. The first lemma makes this notion mathematically precise.

Lemma 3.2 *If $\theta = (1-p\rho)^{-1} > 1$, then*

$$\sum_{w\in\mathcal{A}^k}[\![|S_w(\rho)-1| \le (\rho\delta)^k\theta]\!]\mathbf{P}(w) \ge 1 - \delta^k\theta \tag{32}$$

where $[\![A]\!] = 1$ if A holds, and $[\![A]\!] = 0$ otherwise.

Proof Our proof is the one given in Fayolle and Ward (2005). Note that $S_w(z)-1$ has a term of degree $i \le j$ if and only if $m-i \in \mathcal{P}(w)$ with $1 \le i \le j$. Therefore, for each such i and each $w_1\dots w_i$, there is *exactly one word* $w_{i+1}\dots w_k$ such that $S_w(z)-1$ has a term of degree i. Therefore, for fixed j and k,

$$\begin{aligned}
&\sum_{w\in\mathcal{A}^k}[\![S_w(z)-1 \text{ has a term of degree} \le j]\!]\mathbf{P}(w)\\
&\quad\le \sum_{1\le i\le j}\ \sum_{w_1,\dots,w_i\in\mathcal{A}^i}\mathbf{P}(w_1\dots w_i)\sum_{w_{i+1},\dots,w_k\in\mathcal{A}^{k-i}}[\![S_w(z)-1 \text{ has a term of degree } i]\!]\mathbf{P}(w_{i+1}\dots w_k)\\
&\quad\le \sum_{1\le i\le j}\ \sum_{w_1,\dots,w_i\in\mathcal{A}^i}\mathbf{P}(w_1\dots w_i)p^{k-i} = \sum_{1\le i\le j}p^{k-i} \le \frac{p^{k-j}}{1-p}
\end{aligned} \tag{33}$$

We use $j = \lfloor k/2 \rfloor$. Thus $\sum_{w\in\mathcal{A}^k} [\![\text{all terms of } S_w(z) - 1 \text{ have degree} > \lfloor k/2 \rfloor]\!] \mathbf{P}(w) \geq 1 - \delta^k\theta$.

Note that, if all terms of $S_w(z) - 1$ have degree $> \lfloor k/2 \rfloor$, then

$$|S_w(\rho) - 1| \leq \sum_{i>\lfloor k/2 \rfloor} (\rho p)^i = \frac{(\rho p)^{\lfloor k/2 \rfloor + 1}}{1 - \rho p} \leq \frac{(\rho p)^{k/2}}{1 - \rho p} \leq \frac{\rho^k p^{k/2}}{1 - \rho p} = (\rho\delta)^k \theta\,. \tag{34}$$

This completes the proof of the lemma. □

Using this lemma, we can quickly obtain another result that is similar but slightly stronger.

First consider words w such that $|S_w(\rho) - 1| \leq (\rho\delta)^k\theta$. Write $S_w(z) = \sum_{i=0}^{k-1} a_i z^i$ and $S_{w\alpha}(z) = \sum_{i=0}^{k} b_i z^i$. Observe that either $b_i = 0$ or $b_i = a_i$. The following lemma follows immediately:

Lemma 3.3 *If $\theta = (1 - p\rho)^{-1} + 1$ and $\alpha \in \mathcal{A}$, then*

$$\sum_{w\in\mathcal{A}^k} [\![\max\{|S_w(\rho) - 1|, |S_{w\alpha}(\rho) - 1|\} \leq (\rho\delta)^k\theta]\!] \mathbf{P}(w) \geq 1 - \delta^k\theta\,. \tag{35}$$

Also, the autocorrelation polynomial is never too small. In fact

Lemma 3.4 *Define $c = 1 - \rho\sqrt{p} > 0$. Then there exists an integer $K \geq 1$ such that, for $|w| \geq K$ and $|z| \leq \rho$ and $|u| \leq \delta^{-1}$,*

$$|S_w(z) - uS_{w\alpha}(z) + u| \geq c\,. \tag{36}$$

Proof The proof consists of considering several cases. The only condition for K is $(1 + \delta^{-1})\frac{(p\rho)^{K/2}}{1-p\rho} \leq c/2$. The analysis is not difficult; all details are presented in Ward (2005). □

3.4 Analytic Continuation

Our goal in this section is to prove the following:

Theorem 3.2 *The generating function $M(z, u)$ can be analytically continued for $|u| \leq \delta^{-1}$ and $|z| < 1$.*

The proof requires several lemmas and observations. We always assume $|u| \leq \delta^{-1}$.

Lemma 3.5 *If $0 < r < 1$, then there exists $C > 0$ and an integer K_1 (both depending on r) such that*

$$|D_w(z) - u(D_{w\alpha}(z) - (1 - z))| \geq C \tag{37}$$

for $|w| \geq K_1$ and $|z| \leq r$ (and, as before, $|u| \leq \delta^{-1}$).

Proof Consider the K and c defined in Lemma 3.4, which tells us that, for all $|w| \geq K$, we have

$$|S_w(z) - uS_{w\alpha}(z) + u| \geq c \tag{38}$$

for $|z| \leq \rho$. So, for $|w| \geq K$, we have $|D_w(z) - u(D_{w\alpha}(z) - (1 - z))| \geq (1 - r)c - r^m p^m(1 - \delta^{-1}rp)$. Note that $r^m p^m(1 - \delta^{-1}rp) \to 0$ as $m \to \infty$. Therefore, replacing K by a larger K_1 if necessary, we can without loss of generality assume that $r^m p^m(1 - \delta^{-1}rp) \leq (1 - r)c/2$. So we define $C = (1 - r)c/2$, and the result follows immediately. □

Now we can strengthen the previous lemma by dropping the "K_1", i.e., by not requiring w to be a long word:

Lemma 3.6 *If $0 < r < 1$, then there exists $C > 0$ (depending on r) such that*

$$|D_w(z) - u(D_{w\alpha}(z) - (1 - z))| \geq C \tag{39}$$

for $|z| \leq r$ (and, as before, $|u| \leq \delta^{-1}$).

Proof Consider the K_1 defined in Lemma 3.5. Let C_0 denote the "C" from Lemma 3.5. There are only finitely many w's with $|w| < K_1$, say $w_1, \ldots, w_i$. For each such w_j (with $1 \leq j \leq i$), we note that $D_{w_j}(z) - u(D_{w_j\alpha}(z) - (1 - z)) \neq 0$ for $|z| \leq r$ and $|u| \leq \delta^{-1}$, so there exists $C_j > 0$ such that $|D_{w_j}(z) - u(D_{w_j\alpha}(z) - (1 - z))| \geq C_j$ for all $|z| \leq r$ and $|u| \leq \delta^{-1}$. Finally, we define $C = \min\{C_0, C_1, \ldots, C_i\}$. □

Finally, we prove Theorem 3.2.

Proof Consider $|z| \le r < 1$. We proved in Lemma 3.6 there exists $C > 0$ depending on r such that, for all $|u| \le \delta^{-1}$, we have $\frac{1}{|D_w(z)-u(D_{w\alpha}(z)-(1-z))|} \le \frac{1}{C}$. Setting $u = 0$, we also have $\frac{1}{|D_w(z)|} \le \frac{1}{C}$. Thus

$$|M(z,u)| \le \frac{\mathbf{P}(\beta)\delta^{-1}}{C^2} \sum_{\alpha\in\mathcal{A}} \sum_{w\in\mathcal{A}^*} \mathbf{P}(w)|D_{w\alpha}(z) - (1-z)|\,. \tag{40}$$

Now we use Lemma 3.3. Consider w and α with $\max\{|S_w(\rho)-1|, |S_{w\alpha}(\rho)-1|\} \le (\rho\delta)^m\theta$. It follows immediately that

$$|D_{w\alpha}(z)-(1-z)| = |(1-z)(S_{w\alpha}(z)-1)+z^{m+1}\mathbf{P}(w)\mathbf{P}(\alpha)| \le (1+r)(\rho\delta)^m\theta + r^{m+1}p^m p = O(s^m)\,, \tag{41}$$

where $s = \max\{\rho\delta, rp\}$. Now consider the other w's and α's. We have

$$|D_{w\alpha}(z)-(1-z)| = |(1-z)(S_{w\alpha}(z)-1)+z^{m+1}\mathbf{P}(w)\mathbf{P}(\alpha)| \le \frac{(1+r)p\rho}{1-p\rho} + r^{m+1}p^m p \le \frac{(1+r)p\rho}{1-p\rho}+1\,, \tag{42}$$

so we define $C_1 = \frac{(1+r)p\rho}{1-p\rho} + 1$ to be a value which only depends on r (recall that r is fixed here). Thus

$$\begin{aligned} |M(z,u)| &\le \frac{\mathbf{P}(\beta)\delta^{-1}}{C^2} \sum_{\alpha\in\mathcal{A}} \sum_{m\ge 0} \sum_{w\in\mathcal{A}^m} |\mathbf{P}(w)(D_{w\alpha}(z)-(1-z))| \\ &\le \frac{\mathbf{P}(\beta)\delta^{-1}}{C^2} \sum_{\alpha\in\mathcal{A}} \sum_{m\ge 0} |(1-\delta^m\theta)O(s^m) + \delta^m\theta C_1| \le \frac{\mathbf{P}(\beta)\delta^{-1}}{C^2} \sum_{\alpha\in\mathcal{A}} \sum_{m\ge 0} O(s^m) = O(1) \end{aligned} \tag{43}$$

and this completes the proof of the theorem. □

3.5 Singularity Analysis

We first determine (for $|u| \le \delta^{-1}$) the zeroes of $D_w(z) - u(D_{w\alpha}(z) - (1-z))$ and (in particular) the zeroes of $D_w(z)$.

Lemma 3.7 *There exists an integer $K_2 \ge 1$ such that, for u fixed (with $|u| \le \delta^{-1}$) and $|w| \ge K_2$, there is exactly one root of $D_w(z) - u(D_{w\alpha}(z) - (1-z))$ in the closed disk $\{z \mid |z| \le \rho\}$.*

Proof Let K and c be defined as in Lemma 3.4. Without loss of generality (replacing K by a larger K_2, if necessary), we can also assume that $2(p\rho)^{K_2} < c(\rho-1)$ and $K_2 \ge K_1$ (where K_1 is defined in Lemma 3.5). Also, we can choose K_2 large enough (for use later) such that $\exists c_2 > 0$ with

$$\rho(1-p^{K_2}(1+\delta^{-1}p)) - 1 > c_2 \qquad \text{and thus} \qquad \rho(1-p^{K_2}) - 1 > c_2\,. \tag{44}$$

We recall $0 < p\rho\delta^{-1} < 1$, and thus $0 < 1 - p\rho\delta^{-1} < 1$. Since $|u| < \delta^{-1}$ and $|z| \le \rho$, then for $|w| \ge K_2$ we have $|\mathbf{P}(w)z^m(1-uz\mathbf{P}(\alpha))| \le (p\rho)^m(1+\delta^{-1}\rho p) \le 2(p\rho)^m < c(\rho-1) \le |(S_w(z) - uS_{w\alpha}(z)+u)(\rho-1)|$. Therefore, for z on the circle $\{z \mid |z| = \rho\}$, we have $|\mathbf{P}(w)z^m(1-uz\mathbf{P}(\alpha))| < |(S_w(z) - uS_{w\alpha}(z)+u)(z-1)|$. Equivalently,

$$|(D_w(z) - u(D_{w\alpha}(z)-(1-z))) - ((S_w(z)-uS_{w\alpha}(z)+u)(z-1))| < |(S_w(z)-uS_{w\alpha}(z)+u)(z-1)|\,. \tag{45}$$

Therefore, by Rouché's Theorem, $D_w(z) - u(D_{w\alpha}(z)-(1-z))$ and $(S_w(z) - uS_{w\alpha}(z)+u)(z-1)$ have the same number of zeroes inside the disk $\{z \mid |z| \le \rho\}$. Since $|S_w(z) - uS_{w\alpha}(z) + u| \ge c$ inside this disk, we conclude that $(S_w(z)-uS_{w\alpha}(z)+u)(z-1)$ has exactly one root in the disk. It follows that $D_w(z) - u(D_{w\alpha}(z)-(1-z))$ also has exactly one root in the disk. □

When $u = 0$, this lemma implies (for $|w| \ge K_2$) that $D_w(z)$ has exactly one root in the disk $\{z \mid |z| \le \rho\}$. Let A_w denote this root, and let $B_w = D'_w(A_w)$. Also let $C_w(u)$ denote the root of $D_w(z) - u(D_{w\alpha}(z) - (1-z))$ in the closed disk $\{z \mid |z| \le \rho\}$. Finally, we define

$$E_w(u) := \left(\frac{d}{dz}\left(D_w(z) - u(D_{w\alpha}(z)-(1-z))\right)\right)\bigg|_{z=C_w} = D'_w(C_w) - u(D'_{w\alpha}(C_w)+1)\,. \tag{46}$$

We have precisely determined the singularities of $M(z,u)$. Next, we make a comparison of $M(z,u)$ to $M^I(z,u)$, in order to show that M_n and M^I_n have asymptotically similar behaviors.

3.6 Comparing Suffix Trees to Tries

Now we define

$$Q(z,u) = M(z,u) - M^I(z,u)\,. \tag{47}$$

Using the notation from (11) and (31), if we write

$$\begin{aligned} M^I_{w,\alpha}(z,u) &= \frac{u\mathbf{P}(\beta)\mathbf{P}(w)}{1-z(1-\mathbf{P}(w))}\frac{z\mathbf{P}(w)\mathbf{P}(\alpha)}{1-z(1+u\mathbf{P}(w)\mathbf{P}(\alpha)-\mathbf{P}(w))} \\ M_{w,\alpha}(z,u) &= \frac{u\mathbf{P}(\beta)\mathbf{P}(w)}{D_w(z)}\frac{D_{w\alpha}(z)-(1-z)}{D_w(z)-u(D_{w\alpha}(z)-(1-z))} \end{aligned} \tag{48}$$

then we have proven that

$$Q(z,u) = \sum_{\substack{w\in\mathcal{A}^*\\ \alpha\in\mathcal{A}}} \left(M_{w,\alpha}(z,u) - M^I_{w,\alpha}(z,u)\right). \tag{49}$$

We also define $Q_n(u) = [z^n]Q(z,u)$. We denote the contribution to $Q_n(u)$ from a specific w and α as $Q_n^{(w,\alpha)}(u) = [z^n](M_{w,\alpha}(z,u) - M^I_{w,\alpha}(z,u))$. Then we observe that

$$Q_n^{(w,\alpha)}(u) = \frac{1}{2\pi i}\oint (M_{w,\alpha}(z,u) - M^I_{w,\alpha}(z,u))\frac{dz}{z^{n+1}} \tag{50}$$

where the path of integration is a circle about the origin with counterclockwise orientation.

We define

$$I_{w,\alpha}(\rho,u) = \frac{1}{2\pi i}\int_{|z|=\rho} (M_{w,\alpha}(z,u) - M^I_{w,\alpha}(z,u))\frac{dz}{z^{n+1}}\,. \tag{51}$$

By Cauchy's theorem, we observe that the contribution to $Q_n(u)$ from a specific w and α is exactly

$$\begin{aligned} Q_n^{(w,\alpha)}(u) &= I_{w,\alpha}(\rho,u) - \operatorname*{Res}_{z=A_w}\frac{M_{w,\alpha}(z,u)}{z^{n+1}} - \operatorname*{Res}_{z=C_w(u)}\frac{M_{w,\alpha}(z,u)}{z^{n+1}} \\ &\quad + \operatorname*{Res}_{z=1/(1-\mathbf{P}(w))}\frac{M^I_{w,\alpha}(z,u)}{z^{n+1}} + \operatorname*{Res}_{z=1/(1+u\mathbf{P}(w)\mathbf{P}(\alpha)-\mathbf{P}(w))}\frac{M^I_{w,\alpha}(z,u)}{z^{n+1}}\,. \end{aligned} \tag{52}$$

To simplify this expression, note that

$$\begin{aligned} \operatorname*{Res}_{z=A_w}\frac{M_{w,\alpha}(z,u)}{z^{n+1}} &= -\frac{\mathbf{P}(\beta)\mathbf{P}(w)}{B_w}\frac{1}{A_w^{n+1}} \\ \operatorname*{Res}_{z=C_w(u)}\frac{M_{w,\alpha}(z,u)}{z^{n+1}} &= \frac{\mathbf{P}(\beta)\mathbf{P}(w)}{E_w(u)}\frac{1}{C_w(u)^{n+1}} \\ \operatorname*{Res}_{z=1/(1-\mathbf{P}(w))}\frac{M^I_{w,\alpha}(z,u)}{z^{n+1}} &= \mathbf{P}(\beta)\mathbf{P}(w)(1-\mathbf{P}(w))^n \\ \operatorname*{Res}_{z=1/(1+u\mathbf{P}(w)\mathbf{P}(\alpha)-\mathbf{P}(w))}\frac{M^I_{w,\alpha}(z,u)}{z^{n+1}} &= -\mathbf{P}(\beta)\mathbf{P}(w)(1+u\mathbf{P}(w)\mathbf{P}(\alpha)-\mathbf{P}(w))^n \end{aligned} \tag{53}$$

It follows from (52) that

$$\begin{aligned} Q_n^{(w,\alpha)}(u) &= I_{w,\alpha}(\rho,u) + \frac{\mathbf{P}(\beta)\mathbf{P}(w)}{B_w}\frac{1}{A_w^{n+1}} - \frac{\mathbf{P}(\beta)\mathbf{P}(w)}{E_w(u)}\frac{1}{C_w(u)^{n+1}} \\ &\quad + \mathbf{P}(\beta)\mathbf{P}(w)(1-\mathbf{P}(w))^n - \mathbf{P}(\beta)\mathbf{P}(w)(1+u\mathbf{P}(w)\mathbf{P}(\alpha)-\mathbf{P}(w))^n\,. \end{aligned} \tag{54}$$

We next determine the contribution of the $z = A_w$ terms of $M(z,u)$ and the $z = 1/(1-\mathbf{P}(w))$ terms of $M^I(z,u)$ to the difference $Q_n(u) = [z^n](M(z,u) - M^I(z,u))$.

Lemma 3.8 *The "A_w terms" and the "$1/(1-\mathbf{P}(w))$ terms" (for $|w| \geq K_2$) altogether have only $O(n^{-\epsilon})$ contribution to $Q_n(u)$, i.e.,*

$$\sum_{\substack{|w|\geq K_2\\ \alpha\in\mathcal{A}}} \left(-\operatorname*{Res}_{z=A_w}\frac{M_{w,\alpha}(z,u)}{z^{n+1}} + \operatorname*{Res}_{z=1/(1-\mathbf{P}(w))}\frac{M^I_{w,\alpha}(z,u)}{z^{n+1}}\right) = O(n^{-\epsilon})\,, \tag{55}$$

for some $\epsilon > 0$.

Proof We define

$$f_w(x) = \frac{1}{A_w^{x+1} B_w} + (1 - \mathbf{P}(w))^x \tag{56}$$

for x real. So by (53) it suffices to prove that

$$\sum_{\substack{|w| \geq K_2 \\ \alpha \in \mathcal{A}}} \mathbf{P}(\beta)\mathbf{P}(w) f_w(x) = O(x^{-\epsilon}) \,. \tag{57}$$

Note that $\sum_{\substack{|w| \geq K_2 \\ \alpha \in \mathcal{A}}} \mathbf{P}(\beta)\mathbf{P}(w) f_w(x)$ is absolutely convergent for all x. Also $\bar{f}_w(x) = f_w(x) - f_w(0)e^{-x}$ is exponentially decreasing when $x \to +\infty$ and is $O(x)$ when $x \to 0$ (notice that we utilize the $f_w(0)e^{-x}$ term in order to make sure that $\bar{f}_w(x) = O(x)$ when $x \to 0$; this provides a fundamental strip for the Mellin transform in the next step). Therefore, its Mellin transform $\bar{f}_w^*(s) = \int_0^\infty \bar{f}_w(x) x^{s-1}\, dx$ is well-defined for $\Re(s) > -1$ (see Flajolet et al. (1995) and Szpankowski (2001)). We compute

$$\bar{f}_w^*(s) = \Gamma(s) \left(\frac{(\log A_w)^{-s} - 1}{A_w B_w} + (-\log(1 - \mathbf{P}(w)))^{-s} - 1 \right) \tag{58}$$

where Γ denotes the Euler gamma function, and we note that

$$\begin{aligned} (\log A_w)^{-s} &= \left(\frac{\mathbf{P}(w)}{S_w(1)} \right)^{-s} (1 + O(\mathbf{P}(w))) \\ (-\log(1 - \mathbf{P}(w)))^{-s} &= \mathbf{P}(w)^{-s} (1 + O(\mathbf{P}(w))) \end{aligned} \tag{59}$$

Also

$$\begin{aligned} A_w &= 1 + \frac{1}{S_w(1)} \mathbf{P}(w) + O(\mathbf{P}(w)^2) \\ B_w &= -S_w(1) + \left(-\frac{2S_w'(1)}{S_w(1)} + m \right) \mathbf{P}(w) + O(\mathbf{P}(w)^2) \end{aligned} \tag{60}$$

Therefore

$$\frac{1}{A_w B_w} = -\frac{1}{S_w(1)} + O(|w|\mathbf{P}(w)) \tag{61}$$

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

We define $g^*(s) = \sum_{\substack{|w| \geq K_2 \\ \alpha \in \mathcal{A}}} \mathbf{P}(\beta)\mathbf{P}(w) \bar{f}_w^*(s)$. Then we compute

$$g^*(s) = \sum_{\alpha \in \mathcal{A}} \mathbf{P}(\beta) \sum_{|w| \geq K_2} \mathbf{P}(w) \bar{f}_w^*(s) = \sum_{\alpha \in \mathcal{A}} \mathbf{P}(\beta)\Gamma(s) \sum_{m=K_2}^{\infty} \left(\sup\{q^{-\Re(s)}, 1\}\delta \right)^m O(1) \,, \tag{62}$$

where the last equality is true because $1 \geq p^{-\Re(s)} \geq q^{-\Re(s)}$ when $\Re(s)$ is negative, and also because $q^{-\Re(s)} \geq p^{-\Re(s)} \geq 1$ when $\Re(s)$ is positive. We always have $\delta < 1$. Also, there exists $c > 0$ such that $q^{-c}\delta < 1$. Therefore, $g^*(s)$ is analytic in $\Re(s) \in (-1, c)$. Working in this strip, we choose ϵ with $0 < \epsilon < c$. Then we have

$$\sum_{\substack{|w| \geq K_2 \\ \alpha \in \mathcal{A}}} \mathbf{P}(\beta)\mathbf{P}(w) f_w(x) = \frac{1}{2\pi i} \int_{\epsilon - i\infty}^{\epsilon + i\infty} g^*(s) x^{-s}\, ds + \sum_{\substack{|w| \geq K_2 \\ \alpha \in \mathcal{A}}} \mathbf{P}(\beta)\mathbf{P}(w) f_w(0) e^{-x} \,. \tag{63}$$

Majorizing under the integral, we see that the first term is $O(x^{-\epsilon})$ since $g^*(s)$ is analytic in the strip $\Re(s) \in (-1, c)$ (and $-1 < \epsilon < c$). Also, the second term is $O(e^{-x})$. This completes the proof of the lemma. □

Now we bound the contribution to $Q_n(u)$ from the $C_w(u)$ terms of $M(z, u)$ and the $z = 1/(1 + u\mathbf{P}(w)\mathbf{P}(\alpha) - \mathbf{P}(w))$ terms of $M^I(z, u)$.

Lemma 3.9 *The "$C_w(u)$ terms" and the "$1/(1+u\mathbf{P}(w)\mathbf{P}(\alpha)-\mathbf{P}(w))$ terms" (for $|w| \geq K_2$) altogether have only $O(n^{-\epsilon})$ contribution to $Q_n(u)$, for some $\epsilon > 0$. More precisely,*

$$\sum_{\substack{|w|\geq K_2\\ \alpha\in\mathcal{A}}} \left(-\operatorname*{Res}_{z=C_w(u)} \frac{M_{w,\alpha}(z,u)}{z^{n+1}} + \operatorname*{Res}_{z=1/(1+u\mathbf{P}(w)\mathbf{P}(\alpha)-\mathbf{P}(w))} \frac{M^I_{w,\alpha}(z,u)}{z^{n+1}} \right) = O(n^{-\epsilon}) . \quad (64)$$

Proof The proof technique is the same as the one for Lemma 3.8 above. □

Next we prove that the $I_{w,\alpha}(\rho,u)$ terms in (54) have $O(n^{-\epsilon})$ contribution to $Q_n(u)$.

Lemma 3.10 *The "$I_{w,\alpha}(\rho,u)$ terms" (for $|w| \geq K_2$) altogether have only $O(n^{-\epsilon})$ contribution to $Q_n(u)$, for some $\epsilon > 0$. More precisely,*

$$\sum_{\substack{|w|\geq K_2\\ \alpha\in\mathcal{A}}} I_{w,\alpha}(\rho,u) = O(n^{-\epsilon}) . \quad (65)$$

Proof Here we only sketch the proof. A rigorous proof is given in Ward (2005). Recall that

$$\begin{aligned} I_{w,\alpha}(\rho,u) \quad = \quad & \frac{1}{2\pi i}\int_{|z|=\rho} u\mathbf{P}(\beta)\mathbf{P}(w)\biggl(\frac{1}{D_w(z)}\,\frac{D_{w\alpha}(z)-(1-z)}{D_w(z)-u(D_{w\alpha}(z)-(1-z))} \\ & -\frac{1}{1-z(1-\mathbf{P}(w))}\,\frac{z\mathbf{P}(w)\mathbf{P}(\alpha)}{1-z(1+u\mathbf{P}(w)\mathbf{P}(\alpha)-\mathbf{P}(w))}\biggr)\frac{dz}{z^{n+1}} . \end{aligned} \quad (66)$$

By Lemma 3.7, K_2 was selected to be sufficiently large such that $(\rho p)^m(1-\delta^{-1}\rho p) \leq (\rho-1)c/2$. Thus, writing $C_1 = (\rho-1)c/2$, we have $1/|D_w(z)-u(D_{w\alpha}(z)-(1-z))| \leq 1/C_1$ and thus $1/|D_w(z)| \leq 1/C_1$. Also $1/|1-z(1-\mathbf{P}(w))| \leq 1/c_2$ and $1/|1-z(1+u\mathbf{P}(w)\mathbf{P}(\alpha)-\mathbf{P}(w))| \leq 1/c_2$ by (44). So we obtain

$$|I_{w,\alpha}(\rho,u)| = O(\rho^{-n})\mathbf{P}(w)(S_{w\alpha}(\rho)-1) + O(\rho^{-n})\mathbf{P}(w)O((p\rho)^m) . \quad (67)$$

Thus, by Lemma 3.3, $\sum_{\alpha\in\mathcal{A}}\sum_{|w|=m}|I_{w,\alpha}(\rho,u)| = O(\rho^{-n})O((\rho\delta)^m)$. We conclude $\sum_{\substack{|w|\geq K_2\\ \alpha\in\mathcal{A}}}|I_{w,\alpha}(\rho,u)| = O(\rho^{-n})$, and the lemma follows. □

Finally, we consider the contribution to $Q_n(u)$ from small words $|w|$. Basically, we prove that $|w|$ has a normal distribution with mean $\frac{1}{h}\log n$ and variance $\theta\log n$, where $h = -p\log p - q\log q$ denotes the entropy of the source, and θ is a constant. Therefore, $|w| \leq K_2$ is extremely unlikely, and as a result, the contribution to $Q_n(u)$ from words w with $|w| \leq K_2$ is very small.

Lemma 3.11 *The terms $\sum_{\substack{|w|<K_2\\ \alpha\in\mathcal{A}}}(M_{w,\alpha}(z,u) - M^I_{w,\alpha}(z,u))$ altogether have only $O(n^{-\epsilon})$ contribution to $Q_n(u)$.*

Proof Let D_n denote the depth of the $(n+1)$st insertion in a suffix tree, i.e., $D_n < k$ if and only if

$$X_{n+1}\ldots X_{n+k} \neq X_{i+1}\ldots X_{i+k} \qquad \text{for all } 0 \leq i < n \quad (68)$$

i.e., $D_n = |w|$ in the notation of Section 3.2. Similarly, let D^I_n denote the depth of the $(n+1)$st insertion in a trie built over $n+1$ independent strings, i.e., $D^I_n < k$ if and only if

$$X_1(n+1)\ldots X_k(n+1) \neq X_1(i)\ldots X_k(i) \qquad \text{for all } 1 \leq i \leq n \quad (69)$$

i.e., $D^I_n = |w|$ in the notation of Section 3.1.

Therefore,

$$[z^n]\sum_{\substack{|w|<K_2\\ \alpha\in\mathcal{A}}}(M_{w,\alpha}(z,u)-M^I_{w,\alpha}(z,u)) = \sum_{i<K_2}\sum_{k=1}^{n}\bigl(\mathbf{P}(M_n=k\ \&\ D_n=i) - \mathbf{P}(M^I_n=k\ \&\ D^I_n=i)\bigr)u^k . \quad (70)$$

Noting that $\mathbf{P}(M_n = k\ \&\ D_n = i) \leq \mathbf{P}(D_n = i)$ and $\mathbf{P}(M^I_n = k\ \&\ D^I_n = i) \leq \mathbf{P}(D^I_n = i)$, it follows that

$$[z^n]\sum_{\substack{|w|<K_2\\ \alpha\in\mathcal{A}}}|M_{w,\alpha}(z,u) - M^I_{w,\alpha}(z,u)| \leq \sum_{i<K_2}\sum_{k=1}^{n}\bigl(\mathbf{P}(D_n=i) + \mathbf{P}(D^I_n=i)\bigr)|u|^k . \quad (71)$$

In Jacquet and Szpankowski (1994), the *typical* depth D^T_{n+1} in a trie built over $n+1$ independent strings was shown to be asymptotically normal with mean $\frac{1}{h}\log(n+1)$ and variance $\theta\log(n+1)$. We observe that D^I_n (defined in (69)) and D^T_{n+1} have the same distribution; to see this, observe that $\mathbf{P}(D^I_n < k) = \sum_{|w|=k}\mathbf{P}(w)(1-\mathbf{P}(w))^n = \mathbf{P}(D^T_{n+1} < k)$. Therefore, D^I_n is also asymptotically normal with mean $\frac{1}{h}\log n$ and variance $\theta\log n$. In Ward (2005), we rigorously prove that D^I_n and D_n have asymptotically the same distribution, namely, a normal distribution with mean $\frac{1}{h}\log(n+1)$ and variance $\theta\log(n+1)$. Therefore, considering (71) (and noting that K_2 is a constant), it follows that

$$[z^n]\sum_{\substack{|w|<K_2\\ \alpha\in\mathcal{A}}}|M_{w,\alpha}(z,u)-M^I_{w,\alpha}(z,u)| = O(n^{-\epsilon})\,. \tag{72}$$

This completes the proof of the lemma. □

All contributions to (54) have now been analyzed. We are finally prepared to summarize our results.

3.7 Summary and Conclusion

Combining the last four lemmas, we see that $Q_n(u) = O(n^{-\epsilon})$ uniformly for $|u|\le\delta^{-1}$, where $\delta^{-1} > 1$. For ease of notation, we define $b=\delta^{-1}$. Finally, we apply Cauchy's theorem again. We compute

$$\mathbf{P}(M_n=k)-\mathbf{P}(M^I_n=k) = [u^kz^n]Q(z,u) = [u^k]Q_n(u) = \frac{1}{2\pi i}\int_{|u|=b}\frac{Q_n(u)}{u^{k+1}}\,du\,. \tag{73}$$

Since $Q_n(u)=O(n^{-\epsilon})$, it follows that

$$|\mathbf{P}(M_n=k)-\mathbf{P}(M^I_n=k)| \le \frac{1}{|2\pi i|}(2\pi b)\frac{O(n^{-\epsilon})}{b^{k+1}} = O(n^{-\epsilon}b^{-k})\,. \tag{74}$$

So Theorem 2.1 holds. It follows that M_n and M^I_n have asymptotically the same distribution, and therefore M_n and M^I_n asymptotically have the same factorial moments. The main result of Ward and Szpankowski (2004) gives the asymptotic distribution and factorial moments of M^I_n. As a result, Theorem 2.2 follows immediately. Therefore, M_n follows the logarithmic series distribution, i.e., $\mathbf{P}(M_n=j)=\frac{p^jq+q^jp}{jh}$ (plus some small fluctuations if $\ln p/\ln q$ is rational).

Acknowledgements

The second author warmly thanks Philippe Flajolet for his hospitality during his recent stay at INRIA.

References

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. Proceedings of a colloquium organized by TU Wien.

J. Fayolle and M. D. Ward. Analysis of the average depth in a suffix tree under a markov model. In *International Conference on the Analysis of Algorithms*, Barcelona, 2005.

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

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

P. Jacquet and M. Régnier. Limiting distributions for trie parameters. In *Lecture Notes in Computer Science*, volume 214, pages 196–210. Springer, 1986.

P. Jacquet and M. Régnier. Normal limiting distributions of the size of tries. In *Proc. Performance '87*, pages 209–223, Amsterdam, 1987. North Holland.

P. Jacquet and W. Szpankowski. Analysis of digital tries with markovian dependency. *IEEE Transactions on Information Theory*, 37:1470–1475, 1991.

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.

P. Kirschenhofer and H. Prodinger. Further results on digital search tries. *Theoretical Computer Science*, 58:143–154, 1988.

P. Kirschenhofer, H. Prodinger, and W. Szpankowski. On the variance of the external path in a symmetric digital trie. *Discrete Applied Mathematics*, 25:129–143, 1989.

S. Lonardi and W. Szpankowski. Joint source–channel LZ'77 coding. In *IEEE Data Compression Conference*, pages 273–282, Snowbird, 2003.

M. Lothaire, editor. *Applied Combinatorics on Words*, chapter 7. Cambridge, 2005. Analytic Approach to Pattern Matching by P. Jacquet and W. Szpankowski.

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

W. Szpankowski. A generalized suffix tree and its (un)expected asymptotic behaviors. *SIAM J. Computing*, 22:1176–1198, 1993.

W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. Addison-Wesley, Wiley, 2001.

M. D. Ward. *Analysis of the Multiplicity Matching Parameter in Suffix Trees*. PhD thesis, Purdue University, West Lafayette, IN, May 2005.

M. D. Ward and W. Szpankowski. Analysis of a randomized selection algorithm motivated by the LZ'77 scheme. In *1st Workshop on Analytic Algorithmics and Combinatorics*, pages 153–160, New Orleans, 2004.

J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23:337–343, 1977.