On the Number of 2-Protected Nodes in Tries and Suffix Trees

Jeffrey Gaither, Yushi Homma, Mark Sellke, Mark Daniel Ward

On the Number of 2-Protected Nodes in Tries and Suffix Trees

Jeffrey Gaither[1†], Yushi Homma[1‡], Mark Sellke[1§] and Mark Daniel Ward[2¶]

[1]*Dept. of Mathematics, Purdue University, West Lafayette, IN 47907, USA.*
[2]*Dept. of Statistics, Purdue University, West Lafayette, IN 47907, USA.*

We use probabilistic and combinatorial tools on strings to discover the average number of 2-protected nodes in tries and in suffix trees. Our analysis covers both the uniform and non-uniform cases. For instance, in a uniform trie with n leaves, the number of 2-protected nodes is approximately $0.803n$, plus small first-order fluctuations. The 2-protected nodes are an emerging way to distinguish the interior of a tree from the fringe.

Keywords: retrieval trees, suffix trees, Poissonization, Mellin transforms, pattern matching

1 Introduction

A node in a tree is classified as *k-protected* if the distance (measured in edges) from the node to each descendant that is a leaf is *at least* k. For instance, any node that is not a leaf is 1-protected. In this paper, we study 2-protected nodes, namely, those nodes that have distance at least 2 from each leaf in the tree. I.e., 2-protected nodes are neither leaves nor parents of any leaf.

In a recent flurry of papers, several authors have investigated the behavior of 2-protected nodes:

- Cheon and Shapiro [2008] analyzed the average number of 2-protected nodes in unlabeled, ordered (planar) trees. The average portion of 2-protected nodes in such trees approaches $1/6$ as the number of leaves grows arbitrarily large. In $\{0, 1, 2\}$-trees, also known as Unary-Binary or Motzkin trees, the average portion of 2-protected nodes approaches $10/27$ as the number of leaves increases.

- Mansour [2011] studied the number of 2-protected nodes in k-ary trees, i.e., in unlabeled, ordered (planar) trees for which each internal node has exactly k children. He proved that the total number of 2-protected nodes in all k-ary trees with n internal nodes approaches n/k^k, for fixed k, as $n \to \infty$.

†Email: jgaither@math.purdue.edu
‡Email: yushi@indy.rr.com
§Email: msellke@gmail.com
¶Email: mdw@purdue.edu. Supported by NSF Science & Technology Center for Science of Information Grant CCF-0939370.

- Du and Prodinger [2012+] analyzed the average number of 2-protected nodes in digital search trees (DSTs). They studied the case where the branching is unbiased. They utilized q-series to prove that the average number of 2-protected nodes in a DST built over n strings is $(0.307\ldots)(n)$, plus a fluctuating function of $\log n$ with small amplitude (on the order of 10^{-5}).

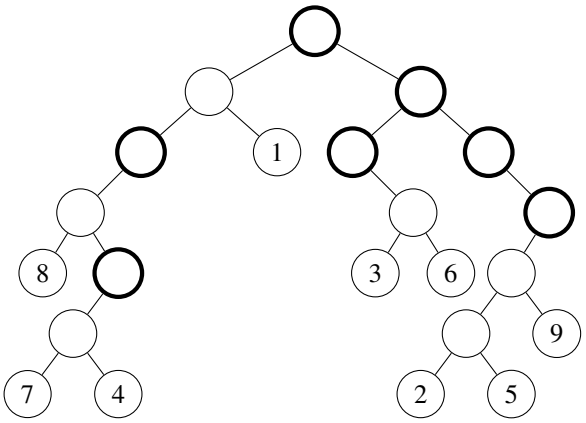
- Mahmoud and Ward [2012+] derived the limiting properties for the number of 2-protected nodes in binary search trees corresponding to permutations grown from uniformly chosen random permutations. They also derived exact expressions for the kth moment of the number of 2-protected nodes in binary search trees, for arbitrarily large k.

We are motivated to study 2-protected nodes because they provide a new method of classification of a node as either: near the fringe (not 2-protected) or away from the fringe of the tree (2-protected).

2 Two-Protected Nodes in Tries

2.1 Definitions

We restrict attention to tries built over binary strings, although we emphasize that the methodology used here is applicable, with only slight generalizations of the ideas, to tries built over strings with letters from a larger alphabet. Let $\mathcal{A} = \{a, b\}$ denote the binary alphabet. When building a trie over a collection of strings, a recursive filtering process occurs, by which each string is placed at the location in the trie corresponding to its shortest unique prefix, as compared to the other strings in the collection. Thus, if $\mathcal{C}$ is a collection of strings with letters from $\mathcal{A}$, then: (1) If $\mathcal{C}$ is empty, then there is no corresponding node in the trie; (2) If $\mathcal{C}$ has one string, then the corresponding node in the trie is a leaf; and (3) If $\mathcal{C}$ has two or more strings, then the splitting process at the analogous node at level j of the tree occurs according to the jth letter of the strings in $\mathcal{C}$.



The trie is built from the strings:

$Y_{1,1}Y_{1,2}Y_{1,3}\ldots = a, b, a, b, a, a, b, a, a\ldots$
$Y_{2,1}Y_{2,2}Y_{2,3}\ldots = b, b, b, a, a, a, b, a, b\ldots$
$Y_{3,1}Y_{3,2}Y_{3,3}\ldots = b, a, b, a, b, b, a, a, b\ldots$
$Y_{4,1}Y_{4,2}Y_{4,3}\ldots = a, a, a, b, a, b, b, b, a\ldots$
$Y_{5,1}Y_{5,2}Y_{5,3}\ldots = b, b, b, a, a, b, a, b, a\ldots$
$Y_{6,1}Y_{6,2}Y_{6,3}\ldots = b, a, b, b, b, b, a, a, a\ldots$
$Y_{7,1}Y_{7,2}Y_{7,3}\ldots = a, a, a, b, a, a, b, b, a\ldots$
$Y_{8,1}Y_{8,2}Y_{8,3}\ldots = a, a, a, a, a, b, b, b, a\ldots$
$Y_{9,1}Y_{9,2}Y_{9,3}\ldots = b, b, b, a, b, a, a, a, a\ldots$

Fig. 1: Example of a trie built from 9 independent strings; 2-protected nodes in bold.

To give an analogous probability model to the splitting process above, we use p and $q := 1 - p$, with $p \geq q$, as the probabilities of a string splitting to the left or right, respectively. In a trie, each string is

generated separately. Moreover, we assume independence within each string, i.e., from letter to letter. Without loss of generality, we define $P(a) = p$ and $P(b) = q$. If the rth string is $Y_{r,1}Y_{r,2}Y_{r,3}\ldots$, and if $(c_1, c_2, c_3, \ldots, c_k)$ is a specific ordered k-tuple with exactly j occurrences of "a" and $k - j$ occurrences of "b", then

$$P(Y_{r,1}Y_{r,2}Y_{r,3}\ldots Y_{r,k} = c_1c_2c_3\ldots c_k) = p^j q^{k-j}. \tag{1}$$

This probability model for the strings induces a unique probability model for the analogous tries built over collections of strings.

Again, a node in a trie is 2-protected if it is neither a leaf nor the parent of a leaf. For example, in Figure 1, there are 9 leaves, 6 internal nodes which are parents of leaves (so *not 2-protected*), and 7 other nodes, which are 2-protected.

Let $T(\mathcal{T}_n^{(I)})$ denote the number of 2-protected nodes in a trie $\mathcal{T}_n^{(I)}$ built over n strings, according to the probability model above. (An "(I)" shows that we are working with 2-protected nodes in tries built over independent strings. Later, we use "(S)" when working with suffix trees.)

Let $X_w(\mathcal{T}_n^{(I)}) = 1$ if the node corresponding to a word w in a trie built over n node is 2-protected; or $X_w(\mathcal{T}_n^{(I)}) = 0$ otherwise. For instance, for the tree $\mathcal{T}_n^{(I)}$ from Figure 1, we have $X_w(\mathcal{T}_n^{(I)}) = 1$, for $w \in \{e, aa, aaab, b, ba, bb, bbb\}$ (where "e" denotes the empty word), and $X_w(\mathcal{T}_n^{(I)}) = 0$ otherwise. Then

$$T(\mathcal{T}_n^{(I)}) = \sum_{w \in \mathcal{A}^*} X_w(\mathcal{T}_n^{(I)}), \tag{2}$$

where $\mathcal{A}^*$ is the collection of finite-length strings, with letters from $\mathcal{A}$. For succinct notation, we define $T_n^{(I)} := T(\mathcal{T}_n^{(I)})$ and $X_{n,w}^{(I)} := X_w(\mathcal{T}_n^{(I)})$.

2.2 Main Results for Tries

We work in a Poissonized model to perform the analysis in Section 3 (see Szpankowski [2001]). Thus, instead of working with tries of fixed size n, we want to sample from a Poisson random variable N_z with average z, and then, conditioned on the value of N_z, we analyze the number of 2-protected nodes in a trie of size N_z. Let $g(z) := \mathbb{E}(T_{N_z}^{(I)})$. We will prove the following:

Theorem 2.1 *Let $g(z)$ be the number of 2-protected nodes in a trie built from a collection of N_z independent strings, where N_z is Poisson with mean z. Let $h = -p\ln p - (1-p)\ln(1-p)$ denote the entropy of the source. Then, for some $\epsilon > 0$, we have*

$$g(z) = \left(\frac{pq+1}{h} - 1\right) z - 1 + \delta(\log z)z + O(z^{-\epsilon}),$$

where δ is a fluctuating function of small magnitude when $\frac{\ln p}{\ln q}$ is rational, and δ converges to 0 otherwise.

Corollary 2.1 *In a trie built from a Poisson number N_z of strings, where the branching to the left-and-right in the tree are equally likely (i.e., $p = q = 1/2$), the number of 2-protected nodes is*

$$g(z) = \left(\frac{5}{4\ln 2} - 1\right) z + \delta(\log z)z - 1 + O(z^{-\epsilon}).$$

Since $\frac{5}{4\ln 2} - 1 = 0.8033688\ldots$, then in a uniform trie with n leaves, there are approximately $0.8033688n$ two-protected nodes, on average, when the trie is built over a set containing a Poisson number of strings. This is smaller than the number of internal nodes in a trie in the uniform case, namely, $n/h \approx 1.44n$, where $h = \ln 2$ is the entropy of a uniform two-letter source.

Corollary 2.2 *Since $g(z) = \mathbb{E}(T^{(I)}_{N_z})$ grows linearly, Theorem 2.1, along with the Depoissonization Theorem 10.4 of Szpankowski [2001], yields that the expected number of nodes, $\mathbb{E}(T^{(I)}_n)$, in a trie built over n independent strings, also has the same asymptotic growth, up to order $O(1)$. In other words,*

$$\mathbb{E}(T^{(I)}_n) = \left(\frac{pq+1}{h} - 1\right) n + \delta(\log n)n + O(1).$$

3 Proofs for Tries

By the linearity of expectation, we have

$$g(z) = \sum_{w \in \mathcal{A}^*} \mathbb{E}(X^{(I)}_{N_z,w}).$$

A key observation is that $X^{(I)}_{N_z,w} = 1$ if and only if two or more of the strings inserted in the trie start with wa, and/or two or more of the strings inserted in the trie start with wb. A Poisson number of strings start with wa, and an independent Poisson number of strings start with wb. Thus

$$\begin{aligned}\mathbb{E}(X^{(I)}_{N_z,w}) =& 1 - P(wa)ze^{-P(wa)z} - P(wb)ze^{-P(wb)z} \\ &+ P(wa)ze^{-P(wa)z}P(wb)ze^{-P(wb)z} - e^{-P(wa)z}e^{-P(wb)z}\end{aligned}$$

We observe $e^{-P(wa)z}e^{-P(wb)z} = e^{-P(w)z}$. Now we consider the Mellin transform

$$g^*(s) := \int_0^\infty g(z) z^{s-1}\, dz$$

of the function $g(z)$. (See Flajolet et al. [1995], Flajolet and Sedgewick [1995, 1996], Szpankowski [2001].) Since $-P(w)z + P(wa)z + P(wb)z = 0$, we can add this term into $\mathbb{E}(X^{(I)}_{N_z,w})$. Notice $1 - P(w)z - e^{-P(w)z}$ and $P(wa)z - P(wa)ze^{-P(wa)z}$ and $P(wb)z - P(wb)ze^{-P(wb)z}$ each have a Mellin strip of $\langle -2, -1\rangle$. Also, $P(wa)ze^{-P(wa)z}P(wb)ze^{-P(wb)z}$ has a Mellin strip of $\langle -2, \infty\rangle$. Thus, the Mellin of $g(z)$ is valid for s in the strip $\langle -2, -1\rangle$, i.e., for s with $-2 < \Re(s) < -1$. Thus, in this strip, we have

$$g^*(s) = \sum_{w \in \mathcal{A}^*} (P(w))^{-s}(pqs(s+1) - p^{-s}s - q^{-s}s - 1)\Gamma(s) = \frac{(pqs(s+1) - p^{-s}s - q^{-s}s - 1)\Gamma(s)}{1 - p^{-s} - q^{-s}}.$$

The pole at $s = -1$ is simple, because the expression in the numerator to the left of $\Gamma(s)$ is 0 at $s = -1$. We retrieve the asymptotics $g(z) = \mathbb{E}(T^{(I)}_{N_z})$ by integrating clockwise around a large rectangle with sides C_1, C_2, C_3, C_4, where C_1 goes from $\frac{-3}{2} - iA$ to $-\frac{3}{2} + iA$; C_2 goes from $-\frac{3}{2} + iA$ to $M + iA$; C_3 goes

from $M + iA$ to $M - iA$; and C_3 goes from $M - iA$ to $\frac{-3}{2} - iA$. The inverse Mellin transform yields

$$\begin{aligned} g(z) &= \frac{1}{2\pi i} \int_{-\frac{3}{2}-i\infty}^{-\frac{3}{2}+i\infty} g^*(s) z^{-s}\, ds \\ &= \lim_{A\to\infty} \sum - \operatorname{Res}[g^*(s)z^{-s}; s = a_\ell] - \lim_{A\to\infty} \frac{1}{2\pi i} \left(\int_{C_2} + \int_{C_3} + \int_{C_4} \right) g^*(s) z^{-s}\, ds, \end{aligned}$$

where the sum is taken over all poles a_ℓ of $g^*(s)z^{-s}$ in the region bounded by $C_1 \cup C_2 \cup C_3 \cup C_4$.

If $\frac{\ln p}{\ln q}$ is rational, say $\frac{\ln p}{\ln q} = r/t$, then the singularities of $g^*(s)z^{-s}$ in the region above are found at $s_\ell = -1 + \frac{2\ell r \pi i}{\ln p}$ and at $s = 0$. As in other studies making comparisons between tries and suffix trees, e.g., Jacquet and Szpankowski [1994, 2005], we have first-order fluctuations in the asymptotic average number of 2-protected nodes when $\frac{\ln p}{\ln q}$ is rational, but no such fluctuations when $\frac{\ln p}{\ln q}$ is irrational. The residue of $g^*(s)z^{-s}$ at $s = -1$ is

$$\operatorname*{Res}_{s=-1} g^*(s) z^{-s} = \left(1 + \frac{p^2 + q^2 - 3}{2h} \right) z,$$

where $h = -p \ln p - (1-p) \ln (1-p)$ is the entropy of the source. The residue of $g^*(s)z^{-s}$ at $s = 0$ is

$$\operatorname*{Res}_{s=0} g^*(s) z^{-s} = 1.$$

The residues from the other singularities only yield small fluctuations, and these fluctuations are only present in the case where $\frac{\ln p}{\ln q}$ is rational. See Flajolet et al. [2010] for more details about the general phenomenon. This completes the proof of Theorem 2.1.

4 Two-Protected Nodes in Suffix Trees

We next make comparisons between the average number of nodes in tries and suffix trees. Our methods continue to be applicable beyond the binary setting, but for conciseness of the presentation, we treat strings built over the alphabet $\mathcal{A} = \{a, b\}$. The letters a and b again correspond to the probabilities p and q in the probability model. We let $\mathcal{S} = X_1 X_2 X_3 \ldots$ denote a sequence of letters drawn independently from $\mathcal{A}$. We take a collection of suffixes from $\mathcal{S}$ by defining the jth suffix as

$$\mathcal{S}_j = X_j X_{j+1} X_{j+2} \ldots.$$

Thus, the jth suffix is the same sequence of characters as S, except that the first $j - 1$ characters are removed. Then we build a trie structure from the collection of the first n suffixes of S, namely, $\mathcal{C}_n := \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n\}$. The entire stochastic structure of the tree depends completely upon the stochastic structure of the string S from which all of the suffixes are drawn. Again, a node in a trie is 2-protected if it is neither a leaf nor the parent of a leaf.

4.1 Main Result for Suffix Trees

Our main result is that, up to a difference of $O(n^\epsilon)$, suffix trees have the same average number of 2-protected nodes as tries, when the probability model (i.e., the values of p and $q = 1 - p$) is the same.

Theorem 4.1 *A suffix tree built from the first n suffixes of a common string has an average $\mathbb{E}(T_n^{(S)})$ of 2-protected nodes, where*

$$\mathbb{E}(T_n^{(S)}) = \left(\frac{pq+1}{h} - 1\right) n + \delta(\log n)n + O(n^\epsilon)$$

where ϵ only needs to satisfy $\epsilon > 1 - \frac{1}{2}\frac{\log(p)}{\log(q)}$, and δ is fluctuating or converging to 0, depending on whether $\frac{\ln p}{\ln q}$ is rational or irrational, respectively.

4.2 Proofs for Suffix Trees

The proof of Theorem 4.1 spans the rest of the paper. Our notation is inherited from Régnier and Denise [2004], Jacquet and Szpankowski [2005]. These methods have early origins in Guibas and Odlyzko [1978, 1981a,b]. For each string $w \in \mathcal{A}^*$, we define

$$\begin{aligned}\mathcal{R}_w &= \{\sigma \mid \sigma \text{ has exactly one occurrence of } w, \text{ which is at the right end}\};\\ \mathcal{U}_w &= \{\sigma \mid w\sigma \text{ has exactly one occurrence of } w, \text{ which is at the left end}\}.\end{aligned}$$

As in Régnier and Denise [2004], we now consider languages that depend on two words. We define

$$\begin{aligned}\widetilde{\mathcal{R}}_{wa} &= \{\sigma \mid \sigma \text{ has exactly one } wa, \text{ which is at the right end, and has no } wb\text{'s}\};\\ \mathcal{M}_{wa,wb} &= \{\sigma \mid wa\sigma \text{ has exactly one } wa, \text{ which is at the left end,}\\ &\qquad\qquad \text{and exactly one } wb, \text{ which is at the right end}\}.\\ \widetilde{\mathcal{U}}_{wa} &= \{\sigma \mid wa\sigma \text{ has exactly one } wa, \text{ which is at the left end, and has no } wb\text{'s}\}.\end{aligned}$$

For each language $\mathcal{L}$, we define $L(z) = \sum_\sigma P(\sigma) z^{|\sigma|}$ as the analogous generating function, where $P(\sigma)$ is the binomial probability defined as in (1).

We define the autocorrelation set $\mathcal{A}_w$ of a word w as the set of all strings σ that are both a prefix and a suffix of w. Then $S_w(z) = \sum_{\sigma\in\mathcal{A}_w} P(\sigma) z^{|\sigma|}$ is the autocorrelation polynomial of w. Similarly, for any two words w and v of the same length, we define $\mathcal{A}_{w,v}$ as the set of all strings σ that are both a prefix of v and a suffix of w. Then $S_{w,v}(z) = \sum_{\sigma\in\mathcal{A}_{w,v}} P(\sigma) z^{|\sigma|}$ is the correlation polynomial of the ordered pair w, v. We note that $S_{w,w}(z) = S_w(z)$, i.e., the correlation polynomial of w with itself is equal to the autocorrelation polynomial of w. For any word w, we define a matrix $\mathbb{D}_w(z)$ that will play a fundamental role in the analysis that follows. We define

$$\mathbb{D}_w(z) := (1-z)\begin{bmatrix} S_{wa,wa}(z) & S_{wa,wb}(z) \\ S_{wb,wa}(z) & S_{wb,wb}(z)\end{bmatrix} + \begin{bmatrix} P(wa)z^{|wa|} & P(wb)z^{|wb|} \\ P(wa)z^{|wa|} & P(wb)z^{|wb|}\end{bmatrix}.$$

4.3 Generating Function for Average Number of 2-Protected Nodes in a Suffix Tree

Each node in a suffix tree corresponds naturally to a string w that describes the path from the root of the suffix tree to the node. To determine when w's node in the suffix tree will be 2-protected, we instead consider the complement, i.e., when would the node corresponding to w fail to be 2-protected? This happens when

- neither wa nor wb occurs within the first $n + (|w| + 1) - 1 = n + |w|$ characters, or
- exactly one copy of wa occurs within the first $n + (|w| + 1) - 1 = n + |w|$ characters, or
- exactly one copy of wb occurs within the first $n + (|w| + 1) - 1 = n + |w|$ characters.

The second and third conditions have one possible overlap, i.e., that there are exactly one copy of wa and exactly one copy of wb. The first condition can be simplified, i.e., there is no occurrence of w within the first $n + |w| - 1$ characters. So, the probability that the node corresponding to w is 2-protected is:

$$\begin{aligned} P(w \text{ is 2-protected}) = 1 &- [z^{n+|w|-1}] \left(\frac{1}{1-z} - R_w(z) \frac{1}{1-z} \right) \\ &- [z^{n+|w|}] \left(R_{wa}(z) U_{wa}(z) + R_{wb}(z) U_{wb}(z) \right) \\ &+ [z^{n+|w|}] \left(\widetilde{R}_{wa}(z) M_{wa,wb}(z) \widetilde{U}_{wb}(z) + \widetilde{R}_{wb}(z) M_{wb,wa}(z) \widetilde{U}_{wa}(z) \right) \end{aligned}$$

Let $T_n^{(S)} = T(\mathcal{T}_n^{(S)})$ be the number of 2-protected nodes in a suffix tree $\mathcal{T}_n^{(S)}$, built from the first n suffixes of a randomly-generated string. The expected value of $T_n^{(S)}$ is found by summing over all w's, i.e.,

$$\begin{aligned} \mathbb{E}(T_n^{(S)}) = \sum_{w \in \mathcal{A}^*} \Bigg(1 &- [z^{n+|w|-1}] \left(\frac{1}{1-z} - R_w(z) \frac{1}{1-z} \right) - [z^{n+|w|}] \left(R_{wa}(z) U_{wa}(z) + R_{wb}(z) U_{wb}(z) \right) \\ &+ [z^{n+|w|}] \left(\widetilde{R}_{wa}(z) M_{wa,wb}(z) \widetilde{U}_{wb}(z) + \widetilde{R}_{wb}(z) M_{wb,wa}(z) \widetilde{U}_{wa}(z) \right) \Bigg). \end{aligned}$$

Using Jacquet and Szpankowski [2005] (p. 339) and Régnier and Denise [2004] (p. 195, 204), we get

$$\begin{aligned} \mathbb{E}(T_n^{(S)}) = [z^n] \sum_{w \in \mathcal{A}^*} P(w) \Bigg(& \frac{z}{(1-z) D_w(z)} - \frac{pz}{(D_{wa}(z))^2} - \frac{qz}{(D_{wb}(z))^2} \\ &+ (pz(\mathbb{D}_w(z)^{-1})_{1,1} + qz(\mathbb{D}_w(z)^{-1})_{2,1})(-(1-z)(\mathbb{D}_w(z)^{-1})_{1,2})((\mathbb{D}_w(z)^{-1})_{2,1} + (\mathbb{D}_w(z)^{-1})_{2,2}) \\ &+ (pz(\mathbb{D}_w(z)^{-1})_{1,2} + qz(\mathbb{D}_w(z)^{-1})_{2,2})(-(1-z)(\mathbb{D}_w(z)^{-1})_{2,1})((\mathbb{D}_w(z)^{-1})_{1,1} + (\mathbb{D}_w(z)^{-1})_{1,2}) \Bigg). \end{aligned} \tag{3}$$

Referring back to $\mathbb{D}_w$, and using the equations

$$\begin{aligned} \det(\mathbb{D}_w(z)) &= (1-z) D_w(z); \\ S_{wa,wb}(z) &= \frac{P(b)}{P(a)} (S_{wa,wa}(z) - 1), \text{ (and analogously when } wa \text{ and } wb \text{ are flipped)}; \\ S_{wa,wa}(z) + S_{wb,wb}(z) &= S_w(z) + P(w) z^{|w|} + 1; \\ D_w(z) &= D_{wa}(z) + D_{wb}(z) - (1-z); \end{aligned}$$

we can simplify the last two lines of (3) into

$$\frac{P(w) z}{D_w(z)^3} \Big(D_{wa}(z)(q S_{wa}(z) - p S_{wb}(z) + p) + D_{wb}(z)(p S_{wb}(z) - q S_{wa}(z) + q) - (1-z) \Big)$$

Thus, we get a generating function from (3) for $\mathbb{E}(T_n^{(S)})$, that naturally decomposes into three parts:

$$\sum_{n\geq 0} \mathbb{E}(T_n^{(S)}) z^n = h_1^{(S)}(z) - h_2^{(S)}(z) + h_3^{(S)}(z),$$

where

$$h_1^{(S)}(z) := \sum_{w\in\mathcal{A}^*} \frac{P(w)z}{(1-z)D_w(z)},$$
$$h_2^{(S)}(z) := \sum_{w\in\mathcal{A}^*} P(w)\left(\frac{pz}{(D_{wa}(z))^2} + \frac{qz}{(D_{wb}(z))^2}\right) = \sum_{w\in\mathcal{A}^*,|w|\geq 1} \frac{P(w)z}{(D_w(z))^2},$$
$$h_3^{(S)}(z) := \sum_{w\in\mathcal{A}^*} \frac{P(w)z}{D_w(z)^3}\Big(D_{wa}(z)(qS_{wa}(z) - pS_{wb}(z) + p) + D_{wb}(z)(pS_{wb}(z) - qS_{wa}(z) + q) - (1-z)\Big).$$

4.4 (Ordinary) Generating Function for Average Number of 2-Protected Nodes in a Trie

Although we already derived exponential and Poissonized generating functions for the average number of 2-protected nodes in a trie, now we need a version that is an ordinary generating function, so that we can compare to the OGF derived above in the suffix tree case. We again use the notations $T_n^{(I)}$ and $X_{n,w}^{(I)}$ from end of Section 2.1. The probability that, in a trie built over n independently-generated strings, the node corresponding to w is 2-protected is

$$\begin{aligned}\mathbb{E}(X_{n,w}^{(I)}) = 1 &- (1-P(w))^n - nP(wa)(1-P(wa))^{n-1} - nP(wb)(1-P(wb))^{n-1} \\ &+ nP(wa)(n-1)P(wb)(1-P(w))^{n-2}.\end{aligned}$$

Thus the expected value of $T_n^{(I)}$ is found by summing over all w's, i.e.,

$$\begin{aligned}\mathbb{E}(T_n^{(I)}) = \sum_{w\in\mathcal{A}^*} \Big(&1 - (1-P(w))^n - nP(wa)(1-P(wa))^{n-1} - nP(wb)(1-P(wb))^{n-1} \\ &+ nP(wa)(n-1)P(wb)(1-P(w))^{n-2}\Big)\end{aligned}$$

This yields a generating function for $\mathbb{E}(T_n^{(I)})$, that also naturally decomposes into three parts:

$$\sum_{n\geq 0} \mathbb{E}(T_n^{(I)}) z^n = h_1^{(I)}(z) - h_2^{(I)}(z) + h_3^{(I)}(z),$$

where

$$h_1^{(I)}(z) := \sum_{w\in\mathcal{A}^*}\sum_{n\geq 0}(1-(1-P(w))^n)z^n = \sum_{w\in\mathcal{A}^*}\frac{P(w)z}{(1-z)(1-z(1-P(w)))},$$

$$h_2^{(I)}(z) := \sum_{w\in\mathcal{A}^*}\sum_{n\geq 0}\left(nP(wa)(1-P(wa))^{n-1}+nP(wb)(1-P(wb))^{n-1}\right)z^n$$

$$= \sum_{w\in\mathcal{A}^*,|w|\geq 1}\frac{P(w)z}{(1-z(1-P(w)))^2},$$

$$h_3^{(I)}(z) = \sum_{w\in\mathcal{A}^*}\sum_{n\geq 0}nP(wa)(n-1)P(wb)(1-P(w))^{n-2}z^n = \sum_{w\in\mathcal{A}^*}\frac{2P(wa)P(wb)z^2}{(1-z(1-P(w)))^3}.$$

So, finally, it suffices to compare the coefficients of z^n in $h_j^{(S)}(z)-h_j^{(I)}(z)$ in the three cases, $j=1,2,3$.

The next lemma, from Jacquet and Szpankowski [1994, 2005], shows that the autocorrelation polynomial of a word is, with high probability, equal to 1 plus terms of relatively large powers and thus small probabilities. We use the notation $[\![A]\!]=1$ if A occurs, and $[\![A]\!]=0$ otherwise.

Lemma 4.1 *Consider $\theta=(1-p\rho)^{-1}$, $\delta=\sqrt{p}$, and $\rho>1$ with $\rho\delta<1$. Then*

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

Lemma 4.2 *Let $\delta=\sqrt{p}$; use again the $\rho>1$ from Lemma 4.1, i.e., such that $\rho\delta<1$. Consider $D_w(z)=(1-z)S_w(z)+P(w)z^m$, where $S_w(z)$ is the autocorrelation polynomial of w. There exists an integer K such that, for each $|w|\geq K$, the polynomial $D_w(z)$ has exactly one root in the disk $|z|\leq\rho$.*

We often refer to this K throughout the rest of the proofs below. We also let A_w denote this unique root. We write $B_w^{(1)}=D'_w(A_w)$, $B_w^{(2)}=D''_w(A_w)$ and $B_w^{(3)}=D'''_w(A_w)$. We use bootstrapping, to get

$$A_w = 1+\frac{1}{S_w(1)}P(w)+O(|w|P(w)^2)\,,$$

$$B_w^{(1)} = -S_w(1)+\left(|w|-\frac{2S'_w(1)}{S_w(1)}\right)P(w)+O(|w|^2P(w)^2)\,,$$

$$B_w^{(2)} = -2S'_w(1)+\left(-|w|+|w|^2-\frac{3S''(1)}{S_w(1)}\right)P(w)+O(|w|^3P(w)^2) \tag{4}$$

$$B_w^{(3)} = -3S''_w(1)+\left(2|w|-3|w|^2+|w|^3-\frac{4S'''_w(1)}{S_w(1)}\right)P(w)+O(|w|^4P(w)^2) \tag{5}$$

Lemma 4.3 Comparing $h_1^{(S)}$ and $h_1^{(I)}$. *Let $\Delta_n^{(1)}=[z^n](h_1^{(S)}(z)-h_1^{(I)}(z))$. Then $\Delta_n^{(1)}=O(n^\epsilon)$ for any $\epsilon>1-\frac{1}{2}\frac{\log p}{\log q}$.*

Proof. For a fixed string $w \in \mathcal{A}^*$, with $|w| \geq K$, the contribution from w to $\Delta_n^{(1)}$ is

$$\begin{aligned}
&\frac{1}{2\pi i}\int_{|z|=\rho} \frac{P(w)z}{(1-z)}\left(\frac{1}{D_w(z)} - \frac{1}{1-z(1-P(w))}\right)\frac{dz}{z^{n+1}} \\
&\quad - \operatorname*{Res}_{z=A_w} \frac{P(w)z}{(1-z)D_w(z)}\frac{1}{z^{n+1}} + \operatorname*{Res}_{z=1/(1-P(w))} \frac{P(w)z}{(1-z)(1-z(1-P(w)))}\frac{1}{z^{n+1}} \\
&\quad - \operatorname*{Res}_{z=1} \frac{P(w)z}{(1-z)D_w(z)}\frac{1}{z^{n+1}} + \operatorname*{Res}_{z=1} \frac{P(w)z}{(1-z)(1-z(1-P(w)))}\frac{1}{z^{n+1}}.
\end{aligned}$$

For fixed $w \in \mathcal{A}^*$, to bound this integral, we note

$$|1-(1-P(w))z-D_w(z)| = |(1-z)(1-S_w(z))+P(w)z(1-z^{|w|-1})| \leq (1+\rho)(S_w(\rho)-1)+(p\rho)^{|w|}.$$

Thus, the total contribution to $\Delta_n^{(1)}$ from the integrals, summed over all $w \in \mathcal{A}^*$, and using Lemma 4.1, is

$$\begin{aligned}
\left|\sum_{w\in\mathcal{A}^*} \frac{1}{2\pi i}\int_{|z|=\rho} \frac{P(w)z}{(1-z)}\left(\frac{1}{D_w(z)} - \frac{1}{1-z(1-P(w))}\right)\frac{dz}{z^{n+1}}\right| &= O\left(\frac{1}{|2\pi|}(2\pi\rho)\sum_{k\geq 0}(\rho\delta)^k\frac{1}{\rho^{n+1}}\right) \\
&= O(\rho^{-n}).
\end{aligned}$$

The residues at $z = 1$ cancel *perfectly*; no approximation or estimate is needed:

$$- \operatorname*{Res}_{z=1} \frac{P(w)z}{(1-z)D_w(z)}\frac{1}{z^{n+1}} + \operatorname*{Res}_{z=1} \frac{P(w)z}{(1-z)(1-z(1-P(w)))}\frac{1}{z^{n+1}} = 0.$$

Finally, to compare the terms with residues at $z = A_w$ and $z = 1/(1-P(w))$, we have

$$\begin{aligned}
&- \operatorname*{Res}_{z=A_w} \frac{P(w)z}{(1-z)D_w(z)}\frac{1}{z^{n+1}} + \operatorname*{Res}_{z=1/(1-P(w))} \frac{P(w)z}{(1-z)(1-z(1-P(w)))}\frac{1}{z^{n+1}} \\
&= -\frac{P(w)}{(1-A_w)B_w^{(1)}A_w^n} + (1-P(w))^n. \qquad (6)
\end{aligned}$$

Thus, we define

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

so that the difference of the residues in (6) is $f_w(n)$. We also define $\overline{f}_w(x) = f_w(x) - f_w(0)$. Thus, for fixed $k \geq K$, we see $\sum_{w\in\mathcal{A}^k}(f_w(x) - f_w(0))$ decreases exponentially as $x \to \infty$, and is $O(x)$ as $x \to 0$.

Thus the Mellin transform $\sum_{w\in\mathcal{A}^k}\overline{f}^*_w(s)$ of $\sum_{w\in\mathcal{A}^k}(f_w(x)-f_w(0))$ exists for $s>-1$, and we have

$$\begin{aligned}
\sum_{w\in\mathcal{A}^k}\overline{f}^*_w(s) &= \sum_{w\in\mathcal{A}^k}\left(-\frac{P(w)}{B^{(1)}_w(1-A_w)}\int_0^\infty\left(\frac{1}{A^x_w}-1\right)x^{s-1}\,dx+\int_0^\infty((1-P(w))^x-1)x^{s-1}\,dx\right)\\
&= \sum_{w\in\mathcal{A}^k}\left(-\frac{P(w)}{B^{(1)}_w(1-A_w)}\Gamma(s)(\log A_w)^{-s}+\Gamma(s)\left(\log\frac{1}{1-P(w)}\right)^{-s}\right)\\
&= \sum_{w\in\mathcal{A}^k}\left(-\Gamma(s)\left(\frac{P(w)}{S_w(1)}\right)^{-s}(1+O(|w|P(w)))+\Gamma(s)P(w)^{-s}(1+O(P(w)))\right)\\
&= \sum_{w\in\mathcal{A}^k}P(w)^{-s}\Gamma(s)\left(-(1/S_w(1))^{-s}(1+O(|w|P(w)))+(1+O(P(w)))\right)\\
&= \sum_{w\in\mathcal{A}^k}P(w)^{-s-1}\Gamma(s)\left(\frac{P(w)(S_w(1)^{-s}-1)}{S_w(1)^{-s}}\right)O(1)\\
&\le (\sup\{q^{-\Re(s)-1},1\})^k(\rho\delta)^k|s|\Gamma(s)O(1)\,,
\end{aligned}$$

where the last line follows from Lemma 4.1 and from $|S_w(1)^{-s}-1|\le|s|\theta\delta^k$, when $|S_w(1)^{-\Re(s)}-1|\le\theta\delta^k$.

We now take any $c<\frac{\log\rho\delta}{\log q}$. Then this value c necessarily satisfies

$$\sup\{q^{-(c-1)-1},1\}\rho\delta<q^{-\frac{\log\rho\delta}{\log q}}\rho\delta=(\rho\delta)^{-1}(\rho\delta)=1,$$

which in turn implies that $g^*(s)$ is analytic for every s with $\Re(s)\in(-\infty,c-1)$. Then for arbitrary $\lambda<c$ we can take the inverse Mellin transform of $g^*(s)$ along the strip $\Re(s)=\lambda-1$ and thereby conclude that the total contribution of the residues of $w\in\mathcal{A}^k$ over all $k\ge K$ is $O(n^{1-\lambda})$. Furthermore, since λ is an arbitrary quantity $<\frac{\log\rho\delta}{\log q}=\frac{1}{2}\frac{\log p}{\log q}$ and ρ may be chosen as close to 1 as we like, we can justifiably choose any $\lambda<\frac{1}{2}\frac{\log p}{\log q}$. And if we set $\epsilon=1-\lambda$ then ϵ exceeds $1-\frac{1}{2}\frac{\log p}{\log q}$ by as small a quantity as we like. Finally we add in all the contributions of short words w with $|w|<K$ and also the contribution from $\sum_w f_w(0)$; both these bounds are O(1) and therefore completely superfluous. This completes the proof of Lemma 4.3.

Lemma 4.4 Comparing $h^{(S)}_2$ and $h^{(I)}_2$. *Let $\Delta^{(2)}_n=[z^n](h^{(S)}_2(z)-h^{(I)}_2(z))$. Then $\Delta^{(2)}_n=O(n^\epsilon)$ for every $\epsilon>1-\frac{1}{2}\frac{\log p}{\log q}$.*

Proof. For a fixed string $w\in\mathcal{A}^*$, with $|w|\ge K$, the contribution from w to $\Delta^{(2)}_n$ is

$$\begin{aligned}
&\frac{1}{2\pi i}\int_{|z|=\rho}P(w)z\left(\frac{1}{(D_w(z))^2}-\frac{1}{(1-z(1-P(w)))^2}\right)\frac{dz}{z^{n+1}}\\
&\qquad-\operatorname*{Res}_{z=A_w}\frac{P(w)z}{(D_w(z))^2}\frac{1}{z^{n+1}}+\operatorname*{Res}_{z=1/(1-P(w))}\frac{P(w)z}{(1-z(1-P(w)))^2}\frac{1}{z^{n+1}}.
\end{aligned}$$

For fixed $w\in\mathcal{A}^*$, to bound this integral, we note

$$\begin{aligned}
|(1-(1-P(w))z)^2-(D_w(z))^2| &= |(1-(1-P(w))z)-(D_w(z))|\times|(1-(1-P(w))z)+(D_w(z))|\\
&\le((1+\rho)(S_w(\rho)-1)+(p\rho)^{|w|})O(1).
\end{aligned}$$

Thus, the total contribution to $\Delta_n^{(2)}$ from the integrals, summed over all $w \in \mathcal{A}^*$, can again be bounded above by applying Lemma 4.1, and is thus of over $O(\rho^{-n})$.

To compare the terms with residues at $z = A_w$ and $z = 1/(1 - P(w))$, we have

$$\begin{aligned}
&- \operatorname*{Res}_{z=A_w} \frac{P(w)z}{(D_w(z))^2} \frac{1}{z^{n+1}} + \operatorname*{Res}_{z=1/(1-P(w))} \frac{P(w)z}{(1 - z(1 - P(w)))^2} \frac{1}{z^{n+1}} \\
&= \frac{P(w)(nB_w^{(1)} + A_w B_w^{(2)})}{(B_w^{(1)})^3 A_w^{n+1}} - nP(w)(1 - P(w))^{n-1}. \qquad (7)
\end{aligned}$$

Thus, we define

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

so that the difference of the residues in (7) is $f_w(n)$. We already have, for fixed $k \geq K$, the property that $\sum_{w \in \mathcal{A}^k} f_w(x)$ decreases exponentially as $x \to \infty$, and is $O(x)$ as $x \to 0$ (so no adjustment by $f_w(0)$ is needed). Thus the Mellin transform $\sum_{w \in \mathcal{A}^k} f_w^*(s)$ exists for $s > -1$, and we have

$$\begin{aligned}
&\sum_{w \in \mathcal{A}^k} f_w^*(s) \\
&= \sum_{w \in \mathcal{A}^k} \left(\frac{P(w)}{(B_w^{(1)})^2 A_w} \int_0^\infty \frac{1}{A_w^x} x^s \, dx + \frac{P(w)B_w^{(2)}}{(B_w^{(1)})^3} \int_0^\infty \frac{1}{A_w^x} x^{s-1} \, dx - P(w) \int_0^\infty (1 - P(w))^{x-1} x^s \, dx \right) \\
&= \sum_{w \in \mathcal{A}^k} \left(\frac{P(w)}{(B_w^{(1)})^2 A_w} (\log A_w)^{-s-1} \Gamma(s+1) + \frac{P(w)B_w^{(2)}}{(B_w^{(1)})^3} (\log A_w)^{-s} \Gamma(s) \right. \\
&\quad \left. - \frac{P(w)}{1 - P(w)} \Gamma(s+1) \left(\log \frac{1}{1 - P(w)} \right)^{-s-1} \right) \\
&= \sum_{w \in \mathcal{A}^k} \left(P(w) \left(\frac{1}{S_w(1)^2} \right) \left(\frac{P(w)}{S_w(1)} \right)^{-s-1} \Gamma(s+1)(1 + O(|w|P(w))) \right. \\
&\quad - P(w) \left(\frac{1}{S_w(1)^3} \right) \left(-2S_w'(1) \right) \left(\frac{P(w)}{S_w(1)} \right)^{-s} \Gamma(s)(1 + O(|w|^2 P(w))) \\
&\quad \left. - P(w)\Gamma(s+1)(P(w))^{-s-1}(1 + O(P(w))) \right) \\
&= \sum_{w \in \mathcal{A}^k} P(w)^{-s-1} \Gamma(s) \left(sP(w)((S_w(1))^{s-1} - 1) + 2S_w'(1)(S_w(1))^{s-3}(P(w))^2 \right) \\
&\leq (\sup\{p^{-\Re(s)-1}, 1\})^k (\rho\delta)^k \Gamma(s) O(1) \,,
\end{aligned}$$

where the last line again follows immediately from the Lemma 4.1. As before, this establishes that the contribution of these residues to $\Delta_n^{(2)}$, taken over all $w \in \mathcal{A}^*$ (not just $w \in \mathcal{A}^k$ for fixed k) is again $O(n^\epsilon)$ for every $\epsilon > 1 - \frac{1}{2} \frac{\log p}{\log q}$.

Finally, we once again have a bound of order $O(1)$ due to the contribution of the shortest words, i.e., those w with $|w| < K$. This completes the proof of Lemma 4.4.

Lemma 4.5 Comparing $h_3^{(S)}$ and $h_3^{(I)}$. *Let $\Delta_n^{(3)} = [z^n](h_3^{(S)}(z) - h_3^{(I)}(z))$. Then $\Delta_n^{(3)} = O(n^\epsilon)$ for every $\epsilon > 1 - \frac{1}{2}\frac{\log p}{\log q}$.*

Proof: For a fixed string $w \in \mathcal{A}^*$, with $|w| \geq K$, the contribution from w to $\Delta_n^{(3)}$ is

$$\frac{1}{2\pi i}\int_{|z|=\rho} \frac{P(w)}{z^{n+1}}\left(\frac{z m_w(z)}{D_w(z)^3} - \frac{2pqP(w)z^2}{(1-z(1-P(w)))^3}\right)dz \tag{8}$$
$$- \operatorname*{Res}_{z=A_w} \frac{P(w)m_w(z)}{D_w(z)^3 z^n} + \operatorname*{Res}_{z=\frac{1}{1-P(w)}} \frac{2pqP(w)^2}{(1-z(1-P(w)))^3 z^{n-1}}.$$

where $m_w(z) = D_{wa}(z)(qS_{wa}(z) - pS_{wb}(z) + p) + D_{wb}(z)(pS_{wb}(z) - qS_{wa}(z) + q) - (1-z)$. To manage $m_w(z)$ throughout the rest of the proof, we need a lemma.

Lemma 4.6 *Let $m_w(z)$ be defined as above. Then we have*

1. $\sum_{w\in\mathcal{A}^k} |m_w(z)|P(w) = O((\rho\delta)^k)$, *uniformly over all z with $|z| \leq \rho$.*
2. $\sum_{w\in\mathcal{A}^k} |m_w(A_w)| = \sum_{w\in\mathcal{A}^k} pq(S_w(1)+1)P(w) + O((\rho\delta)^k)$.
3. $\sum_{w\in\mathcal{A}^k} |m_w^{(j)}(A_w)|P(w) = O((\rho\delta)^k)$ *for $j = 1, 2$.*

Proof: One can write

$$\begin{aligned} m_w(z) = {} & q(S_{wa}(z)-1)\big((1-z)(S_{wa}(z)+1) - D_{wb}(z)\big) \\ & + p(S_{wb}(z)-1)\big((1-z)(S_{wb}(z)+1) - D_{wa}(z)\big) \\ & + pq(S_w(z) + 1 + P(w)z^k))P(w)z^{k+1}, \end{aligned}$$

and from here the results follow fairly easily from Lemma 4.1, although we must bootstrap to get *2* and *3*. This completes the proof of Lemma 4.6.

Now, to bound the integral-term from 8, we note that by Lemma 4.6 we have

$$\sum_{w\in\mathcal{A}^k}\left|\frac{1}{2\pi i}\int_{|z|=\rho}\frac{P(w)}{z^{n+1}}\frac{z m_w(z)}{D_w(z)^3}\,dz\right| = O((\rho\delta)^k),$$

so the sum of this integral over all $w \in \mathcal{A}^*$ is finite. And since $(1-(1-P(w))z)$ can also be uniformly bounded below and $\sum_{w\in\mathcal{A}^*} P(w)^2 < \infty$, we also have

$$\sum_{w\in\mathcal{A}^*}\left|\frac{1}{2\pi i}\int_{|z|=\rho}\frac{P(w)}{z^{n+1}}\frac{2pqP(w)z^2}{(1-z(1-P(w)))^3}\,dz\right| = O(1).$$

Therefore the integral terms make only a finite contribution to $\Delta_n^{(3)}$.
As for the residues, the trie residue-term is

$$\operatorname*{Res}_{z=1/(1-P(w))}\frac{g_w^{(I)}(z)}{z^{n+1}} = \operatorname*{Res}_{z=1/(1-P(w))}\frac{2pqP(w)^2}{(1-z(1-P(w)))^3 z^{n-1}} = -n(n-1)pqP(w)^2(1-P(w))^{n-2}.$$

However, the quantity $\mathrm{Res}_{z=A_w} \frac{zP(w)m_w(z)}{D_w(z)^3 z^{n+1}}$ is rather complex. To represent it we will have to define some auxiliary notation. We set

$$\alpha_w = \frac{1}{B_w^{(1)}}, \quad \beta_w = \frac{B_w^{(2)}}{2(B_w^{(1)})^2}, \quad \gamma_w = \frac{-2B_w^{(1)}B_w^{(3)} + 3(B_w^{(2)})^2}{6(B_w^{(1)})^3}.$$

(It is easy to verify that these three quantities are all uniformly bounded over all w.) We then have

$$\begin{aligned}\operatorname*{Res}_{z=A_w} \frac{zP(w)m_w(z)}{D_w(z)^3 z^{n+1}} = P(w)\Bigg(\frac{1}{2A_w^n}\Bigg((6\alpha_w\beta_w^2 + 3\alpha_w^2\gamma_w)m_w(A_w) + 6\alpha_w^2\beta_w m_w'(A_w) + \alpha_w^3 m_w''(A_w)\Bigg) \\ + \frac{-n}{A_w^{n+1}}\Bigg(3\alpha_w^2\beta_w m_w(A_w) + \alpha_w^3 m_w'(A_w)\Bigg) + \frac{n(n+1)}{2A_w^{n+2}}\Bigg(\alpha_w^3 m_w(A_w)\Bigg)\Bigg).\end{aligned} \tag{9}$$

Lemma 4.6, together the boundedness of the three Greek letters, implies that

$$\sum_{w\in\mathcal{A}^*} \frac{P(w)}{2A_w^n}\Bigg((6\alpha_w\beta_w^2 + 3\alpha_w^2\gamma_w)m_w(A_w) + 6\alpha_w^2\beta_w m_w'(A_w) + \alpha_w^3 m_w''(A_w)\Bigg)$$

is finite, so we need only consider the two terms on the second line of 9. We therefore define a function $f_w(x)$ such that $f_w(n)$ gives the the difference of the relevant residue-parts at n:

$$\begin{aligned}f_w(x) = P(w)\Bigg(-\frac{x}{A_w^{x+1}}\Bigg(3\alpha_w^2\beta_w m_w(A_w) + \alpha_w^3 m_w'(A_w)\Bigg) + \frac{x(x+1)}{2A_w^{n+2}}\Bigg(\alpha_w^3 m_w(A_w)\Bigg) \\ + x(x-1)pqP(w)(1-P(w))^{x-2}\Bigg).\end{aligned}$$

Clearly $f_w(x)$ is $O(x^1)$ as $x \to 0$ and decreases exponentially as $x \to \infty$, so its Mellin transform $f_w^*(s)$ exists in the strip $\langle -1, \infty\rangle$. This Mellin transform is

$$\begin{aligned}f_w^*(s) = P(w)\Gamma(s+1)\Bigg(\log(A_w)^{-s-1}\Bigg(\frac{-3\alpha_w^2\beta_w m_w(A_w) - \alpha_w^3 m_w'(A_w)}{A_w} + \frac{\alpha_w^3 m_w(A_w)}{2A_w^2}\Bigg) \\ + \log(1-P(w))^{-s-1}\frac{pqP(w)}{(1-P(w))^2}\Bigg) \\ + P(w)\Gamma(s+2)\Bigg(\log(A_w)^{-s-2}\frac{\alpha_w^3 m_w(A_w)}{2A_w^2} - \Bigg(\log(1-P(w))^{-s-2}\frac{pqP(w)}{(1-P(w))^2}\Bigg)\Bigg).\end{aligned}$$

We consider the $\Gamma(s+1)$ and $\Gamma(s+2)$ terms separately. Regarding the $\Gamma(s+1)$ term, it is possible (and easier) to obtain the desired bound without any cancellation between the trie and suffix-tree terms. We calculate

$$\sum_{w\in\mathcal{A}^k} P(w)\Gamma(s+1)\Bigg(\log(A_w)^{-s-1}\Bigg(\frac{-3\alpha_w^2\beta_w m_w(A_w)-\alpha_w^3 m'_w(A_w)}{A_w}+\frac{\alpha_w^3 m_w(A_w)}{2A_w^2}\Bigg)$$
$$+\log(1-P(w))^{-s-1}\frac{pqP(w)}{(1-P(w))^2}\Bigg)$$
$$=O(1)\sum_{w\in\mathcal{A}^k}\Bigg(\Bigg(\frac{P(w)}{S_w(1)}\Bigg)^{-s-1}(|m_w(A_w)|+|m'_w(A_w)|)P(w)+P(w)^{-s-1}P(w)^2\Bigg)\Gamma(s+1)$$
$$=O(1)\sum_{w\in\mathcal{A}^k}\Big((q^{-s-1})^k(|m_w(A_w)|+|m'_w(A_w)|)P(w)+(q^{-s})^kP(w)\Big)\Gamma(s+1)$$
$$=O(1)\Big((q^{-s-1})^kO((\rho\delta)^k)+(q^{-s})^k\Big)\Gamma(s+1).$$

From here we proceed as we did for $\Delta_n^{(1)}$ and $\Delta_n^{(2)}$, and show that the total contribution of this portion of the residue over all $w\in\mathcal{A}^*$ is $O(n^\epsilon)$ (subject to our usual restriction of ϵ).
We now turn our attention to the $\Gamma(s+2)$ term,

$$P(w)\Gamma(s+2)\Bigg(\log(A_w)^{-s-2}\frac{\alpha_w^3 m_w(A_w)}{2A_w^2}-\Bigg(\log(1-P(w))^{-s-2}\frac{pqP(w)}{(1-P(w))^2}\Bigg)\Bigg). \tag{10}$$

By Lemma 4.6 we can take

$$\sum_{w\in\mathcal{A}^*} P(w)\Gamma(s+2)\Bigg(\log(A_w)^{-s-2}\frac{\alpha_w^3 m_w(A_w)}{2A_w^2}\Bigg)$$
$$=\sum_{w\in\mathcal{A}^*} P(w)\Gamma(s+2)\times\Bigg(\log(A_w)^{-s-2}\frac{\alpha_w^3 pq(S_w(1)+1)P(w)}{2A_w^2}\Bigg)+O(n^\epsilon);$$

the calculation is straightforward and follows the same lines as the one above. Making this substitution in 10 leaves us with the term

$$P(w)\Gamma(s+2)\Bigg(\log(A_w)^{-s-2}\frac{\alpha_w^3 pq(S_w(1)+1)P(w)}{2A_w^2}-\log(1-P(w))^{-s-2}\frac{pqP(w)}{(1-P(w))^2}\Bigg).$$

By bootstrapping and Lemma 4.6, we then obtain

$$\sum_{w\in\mathcal{A}^k} P(w)\Gamma(s+2)\bigg(\log(A_w)^{-s-2}\frac{\alpha_w^3 pq(S_w(1)+1)P(w)}{2A_w^2} - \log(1-P(w))^{-s-2}\frac{pqP(w)}{(1-P(w))^2}\bigg)$$

$$= \sum_{w\in\mathcal{A}^k} P(w)\Gamma(s+2)\bigg(\bigg(\frac{P(w)}{S_w(1)}\bigg)^{-s-2}(1+O(P(w)))\bigg(-\frac{pq(S_w(1)+1)}{2S_w(1)^3}P(w)+O(|w|P(w)^2)\bigg)$$

$$+P(w)^{-s-2}(1+O(P(w)))pqP(w)\bigg)$$

$$= \sum_{w\in\mathcal{A}^k} P(w)^{-s-1}\bigg(pq\bigg(-\frac{S_w(1)+1}{2S_w(1)^{-s+1}}+1\bigg)P(w)+O(|w|P(w)^2)\bigg)\Gamma(s+2)$$

$$= \sum_{w\in\mathcal{A}^k} P(w)^{-s-1}\bigg(pq\bigg(\frac{2S_w(1)^{-s+1}-S_w(1)-1}{2S_w(1)^{-s+1}}\bigg)P(w)+O(|w|P(w)^2)\bigg)\Gamma(s+2)$$

$$=O(1)(q^{-s-1})^k|s|(\rho\delta)^k\Gamma(s+2).$$

From here we follow our by-now standard procedure and obtain the result that the total contribution from the third term is $O(n^\epsilon)$ for every $\epsilon > 1 - \frac{1}{2}\frac{\log p}{\log q}$, and we are done.

5 Open Problems

Several open problems remain. E.g., precisely characterize the average number of 2-protected nodes in tries and suffix trees with larger alphabets. Analyze higher moments of the number of 2-protected nodes. (We conjecture that the variance of the number of 2-protected nodes has a different first-order for tries versus suffix trees.) Study the average number of k-protected nodes in tries and suffix trees, for $k > 2$.

Acknowledgements

The authors thank Hosam Mahmoud for his helpful comments on a draft of the manuscript, which improved the presentation of the paper. They also thank the reviewers for their very helpful comments. The fourth author (Ward) thanks his colleagues at the Laboratoire d'Informatique de Paris-Nord (LIPN), especially his hosts Frédérique Bassino and Pierre Nicodème, for their hospitality. The manuscript was submitted during Ward's stay at LIPN.

References

G.-S. Cheon and L. W. Shapiro. Protected points in ordered trees. *Applied Mathematics Letters*, 21: 516–520, 2008.

R. R. Du and H. Prodinger. On protected nodes in digital search trees. *Applied Mathematics Letters*, 2012+. In press.

P. Flajolet and R. Sedgewick. Mellin transforms and asymptotics: finite differences and Rice's integrals. *Theoretical Computer Science*, 144:101–124, 1995.

P. Flajolet and R. Sedgewick. The average case analysis of algorithms: Mellin transform asymptotics. Research Report 2956, INRIA, 1996. 93 pages.

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

P. Flajolet, M. Roux, and B. Vallée. Digital trees and memoryless sources: from arithmetics to analysis. In M. Drmota and B. Gittenberger, editors, *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA '10)*, volume AM of *DMTCS Proceedings*, pages 233–260, 2010.

L. Guibas and A. M. Odlyzko. Maximal prefix-synchronized codes. *SIAM Journal on Applied Mathematics*, 35:401–418, 1978.

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

L. Guibas and A. M. Odlyzko. String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory*, 30A:183–208, 1981b.

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. Jacquet and W. Szpankowski. Analytic approach to pattern matching. In M. Lothaire, editor, *Applied Combinatorics on Words*, chapter 7. Cambridge, 2005. See Lothaire [2005].

M. Lothaire, editor. *Applied Combinatorics on Words*. Cambridge, 2005.

H. M. Mahmoud and M. D. Ward. Asymptotic distribution of two-protected nodes in random binary search trees, 2012+. Submitted.

T. Mansour. Protected points in k-ary trees. *Applied Mathematics Letters*, 24:478–480, 2011.

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

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