The Variance of the Number of 2-Protected Nodes in a Trie

Jeffrey Gaither* Mark Daniel Ward†

Dedicated to the memory of Philippe Flajolet

Abstract

We derive an asymptotic expression for the variance of the number of 2-protected nodes (neither leaves nor parents of leaves) in a binary trie. In an *unbiased* trie on n leaves we find, for example, that the variance is approximately $.934n$ plus small fluctuations (also of order n); but our result covers the general (biased) case as well. Our proof relies on the asymptotic similarities between a trie and its Poissonized counterpart, whose behavior we glean via the Mellin transform and singularity analysis.

Keywords: Analysis of algorithms, Mellin transform, Poissonization, retrieval trees, combinatorics on words.

Mathematical subject classification: 05C05, 60C05, 68W32, 68W40.

1 Introduction

A node in a tree is said to be *k-protected* if its distance from every leaf (measured by the number of edges) is at least k. *Every* node in a tree is 0-protected, for instance, while the 1-protected nodes are precisely those nodes that are not leaves. In this paper we restrict our attention to 2-protected nodes, i.e., nodes which are neither a leaf not a parent of a leaf.

In recent years a substantial body of literature has been published about exact enumeration and/or asymptotic expectation of the number of 2-protected nodes in various types of tree models.

- Cheon and Shapiro [1] proved that the expected proportion of 2-protected nodes in a planar tree approaches 1/6, as the total number of leaves in the tree increases. They also show that the average proportion of 2-protected nodes in Motzkin trees (i.e., those trees in which each node has 0, 1, or 2 children) approaches 10/27; in ternary trees (each node has 0 or 3 children), the average proportion approaches 1/81. They have a general program of analysis that extends—with very little modification—to many similar types of trees.
- Mansour [9] established that, in k-ary trees (i.e., in which nodes always have 0 or k offspring) that have n internal nodes, the average proportion of 2-protected nodes approaches n/k^k, as $n \to \infty$.
- Du and Prodinger [3] deduced that the expected number of 2-protected nodes in an unbiased digital search tree of size n is asymptotically $(n)(.307\ldots)$ plus n times a tiny periodic function of $\log n$ (i.e., this function is bounded, with a small maximum amplitude).
- Mahmoud and Ward [8] derived an expression for the number of 2-protected nodes in binary search trees corresponding to uniformly-chosen permutations. They also calculated exact expressions for the kth moment of the number of 2-protected nodes in a BST, using a method that extends to any nonnegative integer k.
- Gaither, Homma, Sellke and Ward [6] discovered the expected number of 2-protected nodes in both tries and suffix trees. The first-order term proved to be the same in both classes of

*Dept. of Mathematics, Purdue University, 150 N. University St., West Lafayette, IN 47907 USA; jgaither@purdue.edu
†Dept. of Statistics, Purdue University, 150 N. University St., West Lafayette, IN 47907 USA; mdw@purdue.edu

tree structures, and in the uniform case was approximately $(n)(.8034\ldots)$.

1.1 Motivation

R*etrie*val trees—henceforth referred to as tries—are one of the most prolific data structures. They were introduced more than 50 years ago by de la Briandais [2] and Fredkin [5]. The precise analysis of the asymptotic characteristic of trie parameters continues to be a topic of broad interest. See, for instance, the recent survey [10], and the many references contained therein, for a thorough analysis of the profile (number of nodes at a given level) of tries.

The variance of the number of 2-protected nodes is of interest because it allows us to determine (in a forthcoming report) the *asymptotic distribution* of the number of 2-protected nodes in tries. The method of solution is also of interest because we are able to derive exact generating functions for the quantities under consideration. A distributional result, in turn, could lead to results which generalize to other kinds of trees, e.g., to the analogous distribution in suffix trees, but attaining the variance is a crucial aspect of this larger analysis.

Furthermore, our result about the variance is (to the best of our knowledge) among the first such results about the number of 2-protected nodes in any tree model. (Mahmoud and Ward [8] paper is an exception, since it yields a method for all moments of the number of 2-protected nodes in binary search trees.)

Two-protected nodes are an emerging parameter of interest, and they have some practical motivations as well. E.g., in a security model with trie structure, a 2-protected node may be taken to represent an entity that has at least two buffers between itself and a vulnerable point; protection is, in this context, highly desirable. In a social-network setting, however, the reverse is true. The classic social-network tree-paradigm uses nodes to represent users on the site, and uses parent-child relationship to represent the act of recruiting a new user to the network. A 2-protected node therefore can be viewed as one who has recruited in the past (i.e., has children and grandchildren), but has not brought anyone new to the network for awhile (i.e., none of its children are themselves leaves). In the former case, a high variance is a definite danger; in the latter it might be viewed as advantageous.

As one additional motivation, we emphasize that the (more general) concept of k-protected nodes in tree models for $k > 2$ seems to invite new investigations in all tree models, as this parameter has not yet received much attention in the combinatorial or asymptotic analysis literature.

2 Definitions

We work with binary strings, i.e., those with letters from $\mathcal{A} = \{a, b\}$. We use $\mathcal{A}^*$ to denote the set of all strings of finite length (including ε, the "empty string" of length 0).

We use a Bernoulli(p) model, in which the letters within a string are always generated independently, and in which the collection of strings is independent too, i.e., there is no dependence between any collection of strings inserted in a trie. If a string S consists of exactly j occurrences of letter a and k occurrences of letter b, then the inherent probability of a string having prefix S is $P(S) = p^j q^k$. We always consider a finite number n of strings inserted in a trie, and thus, with probability 1, each string will have a prefix of finite length that distinguishes it from the other $n-1$ strings in the collection from which the trie is built. An example collection of strings is shown in a trie structure in Figure 1.

We build a trie $\mathcal{T}_n$ over n strings, say $S_1, \ldots, S_n$, according to a mechanism to be described in Section 2.1. We then define

$$T_n := T(\mathcal{T}_n)$$

to be the number of 2-protected nodes in the trie $\mathcal{T}_n$. Our ultimate quantity of interest is the variance $\mathrm{Var}(T_n)$.

We will do most of our work in a Poissonized model. To this end we define $\mathcal{T}_{N_z}$ to be a trie built on N_z random strings, where N_z is Poisson with parameter z; and then let

$$T_{N_z} := T(\mathcal{T}_{N_z})$$

$S_1 = 0110001101\ldots$ $S_9 = 1011111000\ldots$
$S_2 = 1111110101\ldots$ $S_{10} = 1000001111\ldots$
$S_3 = 0111010000\ldots$ $S_{11} = 0110100111\ldots$
$S_4 = 1100101100\ldots$ $S_{12} = 0001010111\ldots$
$S_5 = 1100010000\ldots$ $S_{13} = 1001111001\ldots$
$S_6 = 0000010010\ldots$ $S_{14} = 0110000001\ldots$
$S_7 = 1100001001\ldots$ $S_{15} = 0000110111\ldots$
$S_8 = 0010111100\ldots$ $S_{16} = 0011011001\ldots$

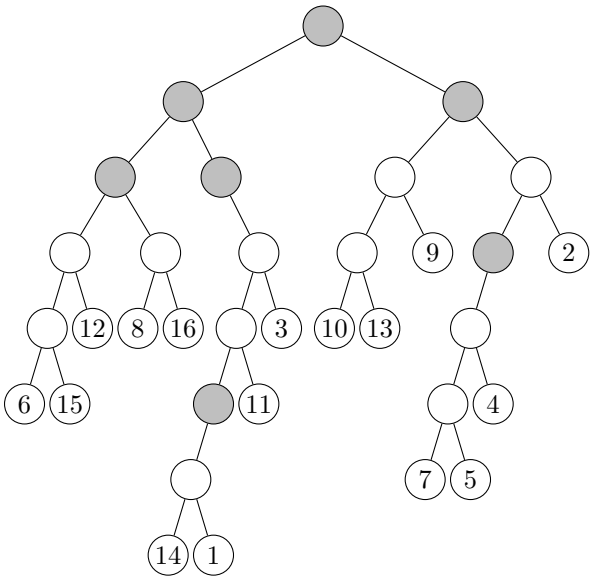

Figure 1: Example of a trie built from 16 randomly-generated strings. All letters are generated independently; each letter of each string is equally likely to be "a" or "b". A branch to the left (respectively, right) at level j indicates that the jth letter of the inserted string is an "a" (respectively, "b"). The 2-protected nodes are shaded in light grey.

be the number of 2-protected nodes in the Poisson-tree $\mathcal{T}_{N_z}$. Finally we define functions for the average, second moment, and variance, all in the Poissonized case:

$$g(z) := \mathbb{E}(T_{N_z}),$$
$$h(z) := \mathbb{E}(T_{N_z}^2),$$
$$v(z) := \mathrm{Var}(T_{N_z}).$$

These three functions are the only ones we will need.

2.1 A Model for Strings in Tries.

Binary tries (and, moreover, tries built over any alphabet) are defined using a recursive scheme of construction. Consider a given set $\mathcal{S}$ of strings, from which a trie will be built. Let $\mathcal{S}\backslash a$ be the subset of $\mathcal{S}$ consisting of strings from $\mathcal{S}$ that began with the character "a", with this leading "a" removed. For instance, if $abbbaab \in \mathcal{S}$, then $bbbaab \in \mathcal{S}\backslash a$. With these definitions, we can now define a trie built on the collection of strings $\mathcal{S}$ as

$$\mathcal{T}(\mathcal{S}) = \begin{cases} \emptyset, & \text{if } \mathcal{S} = \emptyset; \\ Y, & \text{if } \mathcal{S} = \{Y\}; \\ \langle \bullet, \mathcal{T}(\mathcal{S}\backslash a), T(\mathcal{S}\backslash b)\rangle, & \text{otherwise.} \end{cases}$$

The first case is an empty node (does not appear in the trie). The second case is a leaf (i.e., a node where a string is stored in the trie). The third case is of fundamental importance to the construction, because it describes the splitting procedure: a binary trie $\mathcal{T}(\mathcal{S})$ built from a nonempty set $\mathcal{S}$ consists of:

1. a root node $\bullet$;
2. a (possibly empty) left subtree $\mathcal{T}(\mathcal{S}\backslash a)$ consisting of all strings in $\mathcal{S}$ that start with the letter a, with that initial a removed; and
3. a right subtree $\mathcal{T}(\mathcal{S}\backslash b)$ defined analogously.

Therefore, each string is inserted into the trie at the location corresponding to the *shortest unique prefix* of the string. With probability 1, this allows for any finite number of strings to be placed at a finite level in the trie. Figure 1 illustrates the procedure: string

S_3 shares the prefix 011 with two other S_j's, but the prefix 0111 is unique to S_3; thus, S_3 is placed in the trie at the leaf node corresponding to 0111, namely, the shortest distinguishing (i.e., unique) prefix of S_3.

The 2-protected nodes in Figure 1 are drawn in light gray; they are the nodes that are neither leaves nor parents of leaves.

3 Results

Our main result is the following.

Theorem 1 *Let T_n denote the number of 2-protected nodes in a randomly generated binary trie on n leaves, and let the quantity $h := -p\log p - q\log q$ denote the entropy of the source. Then the variance of T_n satisfies*

$$\begin{aligned} Var(T_n) = {} & (c_1 + c_2 - c_3^2)n \\ & + (\delta_1(\log n) - 2c_3\delta_2(\log n) - (\delta_2(\log n))^2)n \\ & + O(n^{1-\epsilon}) \end{aligned}$$

where the constants c_1, c_2 and c_3 are given by

$$\begin{aligned} c_1 = \frac{1}{h}\Bigg(& \frac{2p^3q(2p^2-2pq+5p+3)}{(p+1)^3} \\ & + \frac{2pq^3(2q^2-2pq+5q+3)}{(q+1)^3} \\ & + \frac{pq}{2} - \frac{p^2q^2}{4} - \frac{2p}{p+1} - \frac{2q}{q+1} \\ & + \frac{1}{2} + h - 2pq\left[1 - \frac{p}{(p+1)^2} - \frac{q}{(q+1)^2}\right]\Bigg), \end{aligned}$$

$$c_2 = \frac{2}{h}\sum_{k\geq 2}(-1)^k\frac{(p^k+q^k)^2}{1-p^k-q^k}(pq+1-p^2q^2(k-1)k),$$

$$c_3 = \frac{pq+1}{h} - 1,$$

and δ_1, δ_2 are distinct fluctuating functions of $\log n$ of small ($\sim 10^{-4}$) magnitude when $\frac{\log p}{\log q}$ is rational, and 0 otherwise.

Corollary 1 *If T_n denotes the number of two protected nodes in a trie of n leaves generated from a uniform source (i.e., a source for which $p = q = \frac{1}{2}$), the variance of T_n is asymptotically about $(n)(0.934438705\ldots)$, plus n times small fluctuations.*

Note: We are especially intrigued by the constant c_2, which is defined by an alternating sum; in the uniform case, we have

$$\begin{aligned} c_2 &= \frac{2}{\ln 2}\sum_{k=2}^{\infty}\frac{(-1)^k(k+4)(k-5)}{2^{2k+2}(2^{1-k}-1)} \quad \text{for } p=q=\tfrac{1}{2} \\ &= 1.437275209\ldots \end{aligned}$$

We find this sum very interesting. We are unaware of its appearance elsewhere.

4 Sketch of Proofs

We prove our main result in three lemmas. The first lemma gives us an expression for $\mathrm{Var}(T_n)$, while the second and third are used to provide values that are needed in the first lemma.

Lemma 1 *Let T_n denote the number of 2-protected nodes on a random trie with n leaves. Let $g(z) = \mathbb{E}(T_{N_z})$ and $v(z) = Var(T_{N_z})$ denote (respectively) the expected value and variance of the number of 2-protected nodes in a trie built from N_z leaves, where N_z is Poisson with mean z. Then there exists $\epsilon > 0$ such that*

$$Var(T_n) = v(n) - n(g'(n))^2 + O(n^{1-\epsilon}).$$

Proof. This proof relies on a sharp form of Generalized Depoissonization (see, e.g., Theorem 2 of [7] or Theorem 10.13 of [11]). For our purposes the key result is that if $f(z)$ is the Poisson generating function of a sequence f_n, then one has

$$f_n = f(n) - \frac{n}{2}f''(n) + O(n^{k-2}),$$

provided that $f(z) = O(z^k)$. Gaither et al. [6] showed that the Poissonization $g(z)$ of $\mathbb{E}(T_n)$ is $O(z)$, so $\mathbb{E}(T_n) = g(n) - \frac{n}{2}g''(n) + O(n^{-1})$.

From here the proof is essentially manipulative. Solving for $g(n)$ and squaring, we obtain

$$g(n)^2 = \mathbb{E}(T_n)^2 + ng''(n)\mathbb{E}(T_n) + O(1).$$

Since $g(z)^2 + v(z) = h(z)$, where $h(z) = \mathbb{E}(T_{N_z}^2)$ denotes the second Poissonized moment, then if we add

$$v(n) - \frac{n}{2}\frac{d^2 z}{dz^2}\big(g(z)^2 + v(z)\big)\Big|_{z=n}$$

to both sides, we obtain

$$\begin{aligned}\mathbb{E}(T_n^2) &= \mathbb{E}(T_n)^2 + v(n)\\ &\quad + ng''(n)\big(\mathbb{E}(T_n) - g(n)\big) - n(g'(n))^2 + O(1)\\ &= \mathbb{E}(T_n)^2 + v(n) - n(g'(n))^2 + O(n^{1-\epsilon}).\end{aligned}$$

Subtracting $\mathbb{E}(T_n)^2$ from both sides completes the proof of the lemma. □

The remainder of our proof is devoted to the precise estimation of the quantities $v(n)$ and $g'(n)$.

Lemma 2 *Let $v(z) = Var(T_{N_z})$. Then for some $\epsilon > 0$ we have*

$$v(z) = (c_1 + c_2)z + \delta_1(\log z)z + O(z^{1-\epsilon}),$$

where the definitions of c_1 and c_2 are given in the statement of Theorem 1, and where $\delta_1(\log z)$ is a fluctuating function of small magnitude when $\frac{\log p}{\log q}$ is rational, and 0 otherwise.

Proof. If we let $X_{N_z,w} = 1$ when the node (in the trie) corresponding to w is 2-protected, and $X_{N_z,w} = 0$ otherwise, then we have $T_{N_z} = \sum_{w\in\mathcal{A}^*} X_{N_z,w}$. So we obtain

$$\begin{aligned}v(z) &= \mathrm{Var}(T_{N_z})\\ &= \mathbf{Cov}\Big(\sum_{w\in\mathcal{A}^*} X_{N_z,w}, \sum_{v\in\mathcal{A}^*} X_{N_z,v}\Big)\\ &= \sum_{w,v\in\mathcal{A}^*} \mathbf{Cov}(X_{N_z,w}, X_{N_z,v}).\end{aligned}$$

A crucial observation is that the 2-protectedness of a word w is independent of that of v (and therefore $\mathbf{Cov}(X_{N_z,w}, X_{N_z,v}) = 0$) *unless* w is a prefix of v or v is a prefix of w. There are five ways that this can happen:

1. $v = w$;
2. $v = wax$ for some $x \in \mathcal{A}^*$;
3. $v = wbx$ for some $x \in \mathcal{A}^*$;
4. $w = vax$ for some $x \in \mathcal{A}^*$;
5. $w = vbx$ for some $x \in \mathcal{A}^*$.

In case 1, we have

$$\begin{aligned}\mathbf{Cov}(X_{N_z,w}, X_{N_z,w}) &= \mathbb{E}(X_{N_z,w}^2) - \mathbb{E}(X_{N_z,w})^2\\ &= f_w(z) - f_w(z)^2,\end{aligned}$$

where

$$\begin{aligned}f_w(z) = 1 &- zpP(w)e^{-pP(w)z} - zqP(w)e^{-qP(w)z}\\ &+ z^2pqP(w)^2e^{-P(w)z} - e^{-P(w)z}\end{aligned}$$

is equal to the probability that w is 2-protected in a trie with N_z leaves. (The logic behind this expression is that w is 2-protected if and only if the following two conditions are satisfied: w must appear as the prefix of at least one string inserted in the trie, and (simultaneously) neither wa not wb appears as a prefix of exactly one string inserted in the trie.)

For case 2, we have

$$\begin{aligned}\mathbf{Cov}(X_{N_z,w}, X_{N_z,wax}) &= \mathbb{E}(X_{N_z,w}X_{N_z,wax})\\ &\quad - \mathbb{E}(X_{N_z,w})\mathbb{E}(X_{N_z,wax}).\end{aligned}$$

To analyze $\mathbb{E}(X_{N_z,w}X_{N_z,wax})$, note that if wax is 2-protected, then $X_{N_z,w}$ will be 2-protected if and only if the prefix wb does not appear exactly once as a prefix among strings inserted in the trie. Therefore

$$\mathbb{E}(X_{N_z,w}X_{N_z,wax}) = (1 - zqP(w)e^{-qP(w)z})f_{wax}(z).$$

(It should be evident, at this point, why it is very convenient to know the first letter following w when w is a prefix of v.) We then have

$$\begin{aligned}\mathbf{Cov}(X_{N_z,w},,X_{N_z,wax}) &= f_w(z)f_{wax}(z)\\ &\quad - (1 - zqP(w)e^{-qP(w)z})f_{wax}(z).\end{aligned}$$

The covariances in cases 3–5 can be calculated in analogous ways. We then have, by symmetry,

$$v(z) = v_1(z) + 2v_{2,a}(z) + 2v_{2,b}(z),$$

where

$$v_1(z) := \sum_{w\in\mathcal{A}^*} f_w(z) - f_w(z)^2;$$
$$v_{2,a}(z) := \sum_{w,x\in\mathcal{A}^*} f_{wax}(z)\big(f_w(z) - 1 + zqP(w)e^{-qP(w)z}\big);$$
$$v_{2,b}(z) := \sum_{w,x\in\mathcal{A}^*} f_{wbx}(z)\big(f_w(z) - 1 + zpP(w)e^{-pP(w)z}\big).$$

We will want to take the Mellin transform of $v(z)$, for which we will require a Mellin strip. When one expands $f_w(z)$ out as a Taylor series, the leading term proves to be quadratic; this shows that each $v_j(z) = O(z^2)$ as $z \to 0$. To show that each $v_j(z) = O(z)$ as $z \to \infty$ is more tedious, but this can be seen by first expanding the whole expression, simplifying, and then using calculus to show that, for every $\epsilon > 0$, all remaining terms are uniformly bounded by $Cz^{1+\epsilon}P(w)^{1+\epsilon}$ for some C. So $\langle -2, -1\rangle$ is a valid Mellin strip.

We first take the Mellin transform of $v_1(z)$ in this strip, defined as $v_1^*(s) := \int_0^\infty v_1(z)\, z^{s-1}\, dz$, and we obtain

$$\begin{aligned} v_1^*(s) = m(s)\big[&\big(1-2^{-s}\big)\Gamma(s) \\ &+ \big(p^{-s} + q^{-s} - 2p(p+1)^{-s-1} \\ &\qquad - 2q(q+1)^{-s-1}\big)\Gamma(s+1) \\ &- \big(2^{-s-2}(p^{-s}+q^{-s}) + 3pq + 2^{-s-1}\big)\Gamma(s+2) \\ &+ \big(2pq[p(p+1)^{-s-3} + q(q+1)^{-s-3}]\big)\Gamma(s+3) \\ &+ \big(-p^2q^2 2^{-s-4}\big)\Gamma(s+4)\big] \end{aligned}$$

where

$$m(s) := \sum_{w\in\mathcal{A}^*} P(w)^{-s} = \frac{1}{1-p^{-s}-q^{-s}}.$$

We can now recover $v_1(z)$ via the Inverse Mellin Transform

$$v_1(z) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} z^{-s} v_1^*(s).$$

The real value c can be chosen anywhere in fundamental strip $\langle -1, 0\rangle$. We take $c = -\frac{1}{2}$. Following standard procedure, we then evaluate this integral by "closing the box": that is, we build a box-contour C of four sides, the left of which, C_1, runs from $-\frac{3}{2} - iM$ to $-\frac{3}{2} + iM$ and therefore can be extended to infinity to yield our inverse-Mellin integral. The top and bottom sides C_2 and C_4 run from $-\frac{3}{2}+iM$ to $-1+\epsilon+iM$ and $-1+\epsilon-iM$ to $-\frac{3}{2}-iM$, respectively, for some $\epsilon > 0$. The right-hand side C_3 ranges from $-1+\epsilon+iM$ to $-1+\epsilon-iM$; we will use it to bound our integral.

The integral

$$\lim_{M\to\infty} \frac{1}{2\pi i}\int_{C_1} z^{-s} v_1^*(s)\, ds$$

is precisely our inverse-Mellin integral; the integrals over C_2 and C_4 are $O(e^{-M})$ since the Gamma function decreases exponentially with $|\Im(s)|$, and the integral over C_3 is $O(z^{1-\epsilon})$. And the value of the whole integral around C is then simply the negative sum of the residues contained in C. So

$$v_1(z) = O(z^{1-\epsilon}) - \sum_{s_j\in K} \operatorname*{Res}_{s=s_j} v_1^*(s)$$

where K is the set of poles of v_1^* that lie in the interior of C.

The Γ functions appearing in $v_1^*(s)$ are analytic everywhere within C except at $s = -1$. The function $m(s) = \frac{1}{1-p^{-s}-q^{-s}}$ will have a pole at $s = -1$, and will *also* have infinitely many poles in the strip $\langle -1, -1+\epsilon\rangle$; however, by a result well-explained in Flajolet, Roux, and Vallée's recent survey [4], the total contribution of these poles will be $O(z^{1-\epsilon})$.

Finally, if $\frac{\log p}{\log q}$ is rational, $m(s)$ will have evenly-spaced poles along the line $\Re(s) = -1$, which in total will contribute an oscillating function of $\log z$ of small magnitude. This function will be a sum of terms of form $\alpha_k\Gamma(-1 + 2\pi(k + \frac{a}{\log(p)})i)$, $k \in \mathbb{N}$, where $\frac{\log p}{\log q} = \frac{a}{b} \in \mathbb{Q}$, and $\alpha_k = O(1)$. The "smallness" arises from the Gamma function's exponential rate of decay as its imaginary argument grows large in absolute value. The curious reader is referred to [11] for a more detailed treatment of this point.

Letting $h := -p\log p - q\log q$ denote the entropy of the source, we have

$$\begin{aligned}\operatorname*{Res}_{s=-1} v_1^*(s) = -\frac{1}{h}\Big(&\frac{1}{2} - 2\log 2 + h \\ &+ 2p\log(p+1) + 2q\log(q+1) \\ &- 2pq\Big[1 - \frac{p}{(p+1)^2} - \frac{q}{(q+1)^2}\Big] - \frac{1}{4}p^2q^2\Big).\end{aligned}$$

In the case where $\frac{\log p}{\log q} = \frac{r}{t}$ is rational, we will also have residues at every $z_k = \frac{2\pi i k r}{\log p}$. The net result is a fluctuating function $\bar{\delta}_1$ of small magnitude:

$$\bar{\delta}_1(\log z) = \sum_{k\in\mathbb{Z},k\neq 0} \operatorname*{Res}_{s=z_k} z^{-s} v_1^*(s).$$

Now we want to take the Mellin transform of $v_2(z)$. Unfortunately, $v_2(z)$ is complicated in its given form. Some shifting of the words will greatly simplify the expression. For example, we use

$$\begin{aligned}&\sum_{w\in\mathcal{A}^*} (z^2p^2P(w)^2e^{-2pP(w)z} + z^2q^2P(w)^2e^{-2qP(w)z}) \\ &= \sum_{|w|\geq 1} z^2P(w)^2e^{-2P(w)z}.\end{aligned}$$

In this manner

$$\begin{aligned}\sum_{w,x\in\mathcal{A}^*} &(zpP(w)e^{-pP(w)z}f_{wax}(z) \\ &+ zqP(w)e^{-qP(w)z}f_{wbx}(z))\end{aligned}$$

is seen to be equal to

$$\sum_{|w|\geq 1}\sum_{x\in\mathcal{A}^*} zP(w)e^{-P(w)z}f_{wx}(z).$$

Similarly, we can simplify the following:

$$\begin{aligned}&\sum_{w,x\in\mathcal{A}^*} (1 - z^2pqP(w)^2)e^{-P(w)z}(f_{wax}(z) + f_{wbx}(z)) \\ &= \sum_{w\in\mathcal{A}^*}\sum_{|x|\geq 1} (1 - z^2pqP(w)^2)e^{-P(w)z}f_{wx}(z).\end{aligned}$$

We can then write

$$\begin{aligned}v_2(z) = &\sum_{|w|,|x|\geq 1} (1 + zP(w) - z^2pqP(w)^2)e^{-P(w)z}f_{wx}(z) \\ &+ \sum_{|x|\geq 1} (1 - z^2pq)e^{-z}f_x(z) + \sum_{|w|\geq 1} zP(w)e^{-P(w)z}f_w(z).\end{aligned}$$

Next we convert the fs in the first and second sums into Taylor series and (at last) take Mellin transforms. Dominated convergence allows us to carry the transform inside the sum. Letting $r(s) = (p^{-s}+q^{-s})$, we obtain

$$\begin{aligned}v_2^*(s) = r(s)m(s)&\sum_{|x|\geq 1}\sum_{k\geq 2}\frac{(-P(x))^k}{k!} \\ &\times \big(pq(k)(k-1) + k(p^k+q^k) - 1\big) \\ &\times \big(\Gamma(s+k) + \Gamma(s+k+1) \\ &\qquad - pq\Gamma(s+k+2)\big) \\ &+ \big(1 - s(s+1)pq\big)\sum_{|x|\geq 1}\sum_{k\geq 2}\frac{(-P(x))^k}{k!} \\ &\times (pq(k)(k-1) + k(p^k+q^k) - 1)\Gamma(s+k) \\ &+ m(s)r(s)\big(1 - 2^{-s-1} \\ &- (s+1)\big[(p+1)^{-s-2} + (q+1)^{-s-2}\big] \\ &+ (s+1)(s+2)2^{-s-3}pq\big)\Gamma(s+1).\end{aligned}$$

By shifting, we can evaluate most of this sum explicitly. In the end we find that

$$\begin{aligned}\operatorname*{Res}_{s=-1} v_2^*(s) = -\frac{1}{h}\Big(&\frac{p^3q(2p^2 - 2pq + 5p + 3)}{(p+1)^3} \\ &+ \frac{pq^3(2q^2 - 2pq + 5q + 3)}{(q+1)^3} \\ &+ \frac{pq}{4} - \frac{p}{p+1} - \frac{q}{q+1} \\ &+ \log 2 - p\log(p+1) - q\log(q+1)\Big) \\ &- \frac{1}{h}\sum_{k\geq 2}(-1)^k\frac{(p^k+q^k)^2}{1-p^k-q^k} \\ &\times (pq + 1 - p^2q^2(k-1)k).\end{aligned}$$

When $\frac{\log p}{\log q}$ is rational, we also have periodic residues along $\Re(s) = -1$ which collectively form a small-magnitude function $\bar{\bar{\delta}}_1(\log z)$; since this function is generated by the same poles of the same function as $\bar{\delta}_1(\log z)$, we can combine the two into a single function $\delta_1(\log z)$.

Combining all the residues of $v^*(s) = v_1^*(s) + 2v_2^*(s)$, we find that

$$v(z) = \big(c_1 + c_2\big)z + \delta_1(z) + O(z^{-\epsilon}),$$

where c_1 and c_2 are given in the statement of Theorem 1. This completes the proof of Lemma 2. □

Now we derive an estimate for $g'(z)$, and thus prove our main result.

Lemma 3 *Let $g(z) = \mathbb{E}(T_{N_z})$. Then there is $\epsilon > 0$ such that $g'(z) = c_3 + \delta_2(\log z) + O(z^{-\epsilon})$, where $c_3 = \frac{pq+1}{h} - 1$, and δ_2 is a periodic function of $\log z$ of small magnitude when $\frac{\log p}{\log q}$ is rational, and 0 otherwise.*

Proof. This proof is basically a much easier version of the proof of Lemma 2. We have already noted that the first Taylor-term of $g(z)$ is quadratic; it follows from this that $g'(z) = O(z)$ as $z \to 0$. And we deduce that $g'(z) = O(1)$ as $z \to \infty$ by the same calculus-argument we used in Lemma 2 to show that $v(z)$ was $O(z)$.

We have

$$g(z) = \sum_{w\in\mathcal{A}^*} \big(1 - zpP(w)e^{-pP(w)z} - zqP(w)e^{-qP(w)z} + z^2pqP(w)^2e^{-P(w)z} - e^{-P(w)z}\big),$$

so

$$g'(z) = \sum_{w\in\mathcal{A}^*} \big(P(w)\big(e^{-P(w)z} - pe^{-pP(w)z} - qe^{-qP(w)z}\big) + zP(w)^2\big(p^2e^{-pP(w)z} + q^2e^{-qP(w)z} + 2pqe^{-P(w)z}\big) - z^2P(w)^3pqe^{-P(w)z}\big)$$

The Mellin transform of $g'(z)$ is then

$$\begin{aligned}\frac{dg}{dz}^*(s) &= \sum_{w\in\mathcal{A}^*} P(w)^{-s+1}\Big[\big(1 - p^{-s+1} - q^{-s+1}\big)\Gamma(s) + \big(p^{-s+1} + q^{-s+1} + 2pq\big)\Gamma(s+1) - pq\Gamma(s+2)\Big] \\ &= \frac{1}{1 - p^{-s+1} - q^{-s+1}}\Big[\big(1 - p^{-s+1} - q^{-s+1}\big)\Gamma(s) + \big(p^{-s+1} + q^{-s+1} + 2pq\big)\Gamma(s+1) - pq\Gamma(s+2)\Big].\end{aligned}$$

From here we retrieve $g'(z)$ via the inverse Mellin

$$g'(z) = \frac{1}{2\pi i}\int_{-\frac{1}{2}-i\infty}^{-\frac{1}{2}+i\infty} z^{-s}\frac{dg}{dz}^*(s)\,ds,$$

which we calculate by closing a box whose right-hand side lies on the line $\Re(s) = \epsilon$ for some small ϵ. Calculating the residues, we find

$$\begin{aligned}\operatorname*{Res}_{s=0}\frac{dg}{dz}^*(s) &= -\frac{p\log p + q\log q + pq + 1}{h} \\ &= 1 - \frac{pq+1}{h};\end{aligned}$$

and if $\frac{\log p}{\log q}$ is rational, we sum up the residues along the imaginary axis and obtain

$$\delta_2(\log z) = \sum_{k\in\mathbb{Z},k\neq 1} \operatorname*{Res}_{z=z_k}\frac{dg}{dz}^*(s),$$

a periodic function of $\log z$ of small magnitude. If $\frac{\log p}{\log q}$ is irrational, then we set $\delta_2 = 0$.

We then have

$$g'(z) = \frac{pq+1}{h} - 1 + \delta_2(\log z) + O(z^{-\epsilon}),$$

and squaring $g'(z)$ yields the desired result. □

5 Open Questions

The present inquiry suggests four interesting questions for further investigation.

1. The first order asymptotics of the variance and expectation of the number of 2-protected nodes T_n in a random trie are now known. In a forthcoming report, we will analyze the limiting distribution of T_n. We believe that $\frac{T_n - \mathbb{E}T_n}{\sqrt{\operatorname{Var}(T_n)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$.

2. What can be established about the variance of T_n when $\mathcal{T}_n$ is some species of tree other than a trie? Poissonization lends itself naturally to the analysis of tries. The analysis of the variance of $\mathcal{T}_n$ in other tree structures might lead to interesting results.

3. The sum defining the constant

$$c_2 := \frac{1}{2h}\sum_{k\geq 2}(-1)^k\frac{(p^k+q^k)^2}{1-p^k-q^k}(pq+1-p^2q^2(k-1)k)$$

is not quite like any that we have ever seen before. Are there connections of this constant with the analysis of other trie parameters? We would be pleased to hear about any such connections.

Acknowledgements

The work of M. D. Ward is supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

We thank H. Mahmoud for several helpful conversations about this problem and for discussions about alternative approaches, using recurrence relations.

We also thank H.-K. Hwang and M. Fuchs, who independently alerted us about a small mistake in the error term of [6], which we intend to rectify in a future publication that unifies those results, with the present results, and with an all-new analysis of the limiting distribution of the number of 2-protected nodes in tries.

We are especially thankful for the lucid and informative treatise of P. Flajolet, M. Roux and B. Vallée [4]. It has been very helpful to us in better understanding the singularities in the present problem. It also describes the singularities related to the analysis of tries in a much more general sense (e.g., for strings built from alphabets with more than 2 letters).

Finally, we extend our thanks to M. Sellke and Y. Homma for the benefit of their collaboration in the precursor to this paper [6], which considered the expected number of 2-protected nodes in tries and suffix trees.

References

[1] Gi-Sang Cheon and Louis W. Shapiro. Protected points in ordered trees. *Applied Mathematics Letters*, 21:516–520, 2008.

[2] René de la Briandais. File searching using variable length keys. In *Papers presented at the March 3–5, 1959, Western Joint Computer Conference*, IRE-AIEE-ACM '59 (Western), pages 295–298, New York, 1959.

[3] Rosena R.X. Du and Helmut Prodinger. On protected nodes in digital search trees. *Applied Mathematics Letters*, 2012+. In press.

[4] Philippe Flajolet, Mathieu Roux, and Brigitte Vallée. Digital trees and memoryless sources: from arithmetics to analysis. *Discrete Mathematics and Theroetical Computer Science*, AM:233–260, 2010.

[5] Edward Fredkin. Trie memory. *Communications of the ACM*, 3:490–499, 1960.

[6] Jeffrey Gaither, Yushi Homma, Mark Sellke, and Mark Daniel Ward. On the number of 2-protected nodes in tries and suffix trees. *Discrete Mathematics and Theoretical Computer Science*, AQ:381–398, 2012.

[7] Philippe Jacquet and Wojciech Szpankowski. Analytical depoissonization and its applications. *Theoretical Computer Science*, 201:1–62, 1998.

[8] Hosam M. Mahmoud and Mark Daniel Ward. Asymptotic distribution of two-protected nodes in random binary search trees, 2012+. in press.

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

[10] Gahyun Park, Hsien-Kuei Hwang, Pierre Nicodème, and Wojciech Szpankowski. Profiles of tries. *SIAM Journal on Computing*, 38:1821–1880, 2009.

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