Variance of Size in Regular Graph Tries *

Philippe Jacquet
Bell Labs
Alcatel Lucent
91620 Nozay, France
Email: philippe.jacquet@alcatel-lucent.com

Abram Magner
Dept. Computer Science
Purdue University
W. Lafayette, IN 47907 U.S.A.
Email: anmagner@purdue.edu

Abstract

Graph tries are a generalization of classical digital trees: instead of being built from strings, a G-trie is built from *label functions* on the graph G. In this work, we determine leading order asymptotics for the variance of the size of a G-trie built on a memoryless source on a uniform alphabet distribution, where G is a member of a large class of infinite, M-regular directed, acyclic graphs with $M > 1$ fixed. In particular, this covers the cases of trees and grids. We find that, in such tries, the variance is of order $\Theta(n^{\rho'})$, for some ρ' depending on G which is minimized when G is a tree. We also give an explicit expression for ρ' in the case where G is a grid, with $M = 2$.

Key Words: graph tries, digital trees, Mellin transform, size, variance.

1 Introduction

A *graph trie* [5] is a generalization of the classical *tries*, which are data structures built on words. Given a finite alphabet $\mathcal{A}$ of size A, an infinite directed acyclic graph (DAG) G with a designated root vertex r, and n *label functions* $L_j : V(G) \to \mathcal{A}$, the G-trie GT_n associated with $\{L_j\}_{j=1}^n$ is a tree each of whose internal nodes is a pair consisting of a finite, r-rooted path in G and a labelling of the nodes of the path. The pair $(\mathcal{P}, \omega)$, with $\mathcal{P}$ a path in G starting at r and ω a string over $\mathcal{A}$ of length equal to that of the path, is an internal node in T if and only if there exists a pair $i \neq j$ such that both L_i and L_j label the path $\mathcal{P}$ with the string ω. The root node of GT_n is a special case of this and corresponds to the empty path paired with the empty string. A node $(\mathcal{P}, \omega)$ is a *leaf* of GT_n if and only if there exists exactly one label function L_j which labels $\mathcal{P}$ with ω and removing the last symbol of ω and the last vertex of $\mathcal{P}$ results in a pair which is an internal node of GT_n.

*This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370.

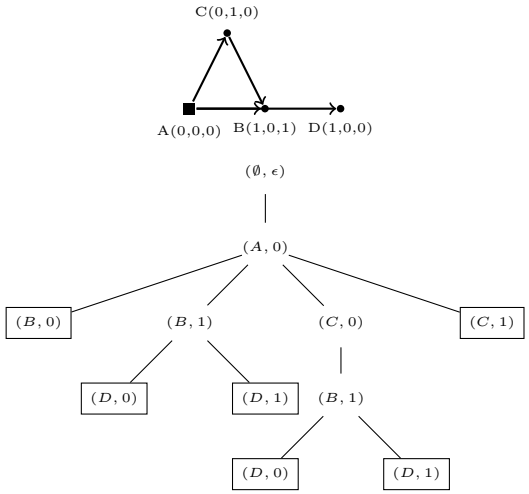

Figure 1: A DAG G with root vertex A and its associated graph trie GT built on three label functions from the alphabet $\mathcal{A} = \{0, 1\}$. The triple on each vertex of G gives the values of the three label functions on that vertex.

Finally, a node $(\mathcal{P}_1, \omega_1)$ is a child of $(\mathcal{P}_2, \omega_2)$ in GT_n if and only if both nodes exist in GT_n and both

- $\mathcal{P}_1$ consists of $\mathcal{P}_2$ with a single node in G appended to the end, and
- ω_1 consists of ω_2 with a single symbol appended to the end.

See Figure 1 for an illustration.

In this paper, we study the asymptotics of the variance of the *size* (i.e., the number of internal nodes, denoted S_n) of a G-trie under a memoryless source model; that is, the label functions are independent and identically distributed, with the individual labels being independently and identically distributed over the alphabet. Whereas [5] considered in its variance analysis only regular trees, we extend the work by

considering arbitrary regular graphs, including trees and rectangular lattices as special cases. This much larger class results in a richer dependence structure among nodes in the induced trie, which leads to a more challenging analysis. Specifically, since two paths in G may meet and diverge several times, two nodes in different subtrees in the trie may be dependent even after conditioning on their common prefix. This phenomenon does not arise when G is a tree.

Analytically, the challenge of the analysis comes from the fact that the Poisson variance contains terms which are not immediately amenable to analysis by Mellin transforms. However, by making appropriate approximations, we reduce the problem to singularity analysis of several double Mellin transforms [7] $V^*(s_1, s_2)$. This function is expressed in terms of a matrix of generating functions $\mathbf{N}(x, y)$, about which we necessarily only have partial information, since it depends on the structure of G. Furthermore, we only have access to a closed form for $\mathbf{N}(x, y)$ in certain special cases. All of this contributes to the difficulty of locating the singularities of $V^*(s_1, s_2)$. Furthermore, after de-Poissonization, we find that a cancellation of terms occurs (as happens in the uniform case when G is a tree), so that we must resort to the *corrected* Poisson variance to estimate one of the lower-order terms.

As for main results, first recall that it was shown in [5] that the expected value of the size S_n is $\Theta(n^\rho)$, where in the uniform case, ρ is given by a formula in terms of M (the out-degree of G) and A (the size of the alphabet):

$$\rho = 1 + \frac{\log M}{\log A},$$

always with A assumed greater than M (called the "non-explosive" condition, since otherwise $\mathbb{E}[S_n]$ explodes to infinity). Here we find that the variance V_n of the size satisfies

$$V_n = \Theta(n^{\rho'}),$$

with $\rho' \geq \rho$ depending on G, and with equality being achieved when G is a tree, as was shown in [5]. That is, the order of growth does not depend on the chosen root node, and it is minimized when G is a tree. We also hint at the (more complicated) analysis in the case where the symbol probabilities are not uniform.

The primary prior work on graph tries is [5], which provided their definition. In addition, the author provided asymptotic expansions for general G for the expected value of the size and insertion cost, as well as for the variance of both parameters in the case where G is a complete M-ary tree (all under a memoryless source model). The same (and many other) parameters have been considered in the context of classical digital trees, including tries (that is, graph tries in which G is a simple sequence of nodes) [8], digital search trees [2], and PATRICIA tries [1].

The rest of the paper is organized as follows: in Section 2, we state the main results, and we give proof sketches in Section 3. We then discuss extending the analysis to the case of a non-uniform distribution over the alphabet in Section 4.

2 Main Results

Here we fix notation, define relevant generating functions, and formally state the main results.

Throughout, we say that a path in a graph G has length ℓ if it consists of ℓ edges (equivalently, $\ell + 1$ vertices). We also need to define the notion of *common nodes* between paths. Let two paths be given by the lists of vertices $a_0, \ldots, a_{\ell_a}$ and $b_0 = a_0, \ldots, b_{\ell_b}$. A node x in G is said to be a *common node* of the two paths if there exist two indices i, j such that $x = a_i = b_j$. We also say that the two paths *meet* at this node a_i. A node x in G and which appears at least once in one of the paths is said to be an *unshared* node if it does not appear in the other path. For example, consider the two paths $\mathcal{P}_1 = (0, 1, 2, 3, 1, 2, 3, 4, 5), \mathcal{P}_2 = (0, 2, 4, 2, 5)$. Then the two paths have 4 common nodes $\{0, 2, 4, 5\}$, $\mathcal{P}_1$ has two unshared nodes: 1 and 3, and $\mathcal{P}_2$ has no unshared nodes.

Let G denote a fixed M-regular infinite directed graph with a designated root vertex r, and let $\mathcal{A}$ be a finite alphabet. We denote by p_α the probability of the symbol $\alpha \in \mathcal{A}$, and, for a string $\omega \in \mathcal{A}^k$, we denote by $P(\omega)$ the probability of ω under a product distribution. The function

$$H(x) = \sum_{\alpha \in \mathcal{A}} p_\alpha^{-x},$$

which is equal to A^{1+x} in the uniform case, will play a role in our analysis, as will the constant ρ such that

$$H(-\rho) = \frac{1}{M}.$$

We are interested in the variance V_n of the size S_n of the G-trie built on n random label functions. We denote by

$$S(z) = \sum_{n \geq 0} \mathbb{E}[S_n] \frac{z^n}{n!} e^{-z}$$

the Poisson transform of the first moment of S_n.

Of particular interest are two classes of M-regular DAGs: *translation-invariant* and *time-expanded*. A graph G is said to be translation-invariant if, for any two vertices v and w, the subgraphs induced by those nodes x for which there is a path from v or w, respectively, to x are isomorphic.

Given a finite DAG G, its *time expansion* G' (first defined in [5]) is defined as follows: to each vertex $v \in V(G)$, we associate a sequence of vertices $v = v_0, v_1, \ldots$ in $V(G')$. Then G' contains an edge from v_i to w_{i+1} if and only if $v = w$ or there is an edge in G from v to w. A graph G' is then said to be *time-expanded* if there exists some G for which G' is the time expansion of G.

We now have enough notation to state our main result.

THEOREM 2.1. *Consider the uniform distribution over $\mathcal{A}$. Throughout, we assume $M < |\mathcal{A}|$.*

For arbitrary M-regular directed, acyclic graphs G, recall that the expected value of the size S_n satisfies

$$\mathbb{E}[S_n] = \Theta(n^\rho).$$

Similarly, when G is an M-regular tree,

$$V_n = \Theta(n^\rho).$$

Again for a large class of M-regular graphs G (in particular, the union of the classes of translation-invariant and time-expanded graphs),

$$V_n = \Theta(n^{\rho'}),$$

where $\rho' \geq \rho$.

In the special case of the 2-dimensional grid (an example of a translation invariant graph), we have an explicit expression for ρ':

$$\rho' = \frac{\log\left(\frac{4}{A^{-1}(2-A^{-1})}\right)}{\log A} = \rho + O\left(\frac{1}{A \log A}\right).$$

The significance of this result is that it confirms our intuition that the polynomial order of the variance of the size is minimized (at least over the set of graphs which we consider here) when G is a tree.

3 Proof Sketches

We give a sketch of the proof of Theorem 2.1. Along the way, we hint at how to generalize the stated result to a class of graphs which is larger still than the specified ones, but we postpone a full analysis to future work.

At a high level, the proof goes as follows: we first derive an exact summation formula for the *Poisson variance*

$$V(z) = Q(z) - (S(z))^2,$$

where $Q(z)$ and $S(z)$ are the Poisson second and first moments of S_n, respectively, of the size, by decomposing the set of pairs of potential trie nodes (i.e., path/label string pairs) into subsets based on the number of overlapping and nonoverlapping vertices in the paths, then on the labels. This formula contains a complicated Hadamard product, for which we develop an explicit expression using the Poisson splitting property. To determine asymptotics of $V(z)$ for $z \to \infty$, the rough plan is to use the Mellin transform and singularity analysis, but not all terms of the expression for $V(z)$ have simple Mellin transforms. To resolve this, we observe that the terms of the summation can be expressed as a function of a quantity β, such that those terms for which $\beta \to 0$ provide the dominant contribution. Thus, we Taylor expand the terms of the summation with respect to β around 0. We then further bound the error term in the expansion.

At this point, we have an asymptotic expression for $V(z)$ which is amenable to analysis by (iterated) Mellin transforms. Each resulting transformed function is written in terms of a vector of generating functions $\mathbf{R}(v, x, y)$ of pairs of paths with a given overlap size (each entry of the vector corresponds to a particular root vertex in G). Then, in order to conduct the singularity analysis, we prove several identities relating $\mathbf{R}(v, x, y)$ to simpler generating functions of pairs of paths. By combinatorial considerations, we are able to determine partial information about the coefficients of these generating functions, which in turn yields information about the locations of their singularities. This allows us to prove that the variance of the size is minimized when G is a tree and, in some special cases, further allows us to explicitly compute the polynomial order of growth of V_n with respect to n.

The final step of the analysis consists of a routine verification of the technical conditions of de-Poissonization in order to transfer asymptotics of $V(n)$ to V_n.

We now move on to the details, starting with defining the combinatorial classes and generating functions which will appear in the analysis.

3.1 Definitions of Relevant Generating Functions

Several sequences of numbers related to paths in G will be important: fix a root vertex r, and let $\mathcal{R}_{l,k,k'}$ denote the set of pairs (P_1, P_2) of paths starting at the root with exactly l common nodes other than the root (i.e., there are nodes $v_1, v_2, \ldots, v_l$, all distinct and not equal to r, such that both P_1 and P_2 contain each v_i, and no other node in the graph besides r is contained in both P_1 and P_2), with P_1 having k (distinct) nodes which are not shared with P_2, and P_2 having k' unshared nodes.

Let $R_{l,k,k'} = |\mathcal{R}_{l,k,k'}|$. Also, let

$$R(v, x, y) = \sum_{l,k,k'=0}^{\infty} R_{l,k,k'} v^l x^k y^{k'}.$$

Now, since we only assume regularity of the underlying graph (instead of the much stronger translation invariance), we form the infinite-dimensional vector $\mathbf{R}(v, x, y)$, whose entries are indexed by the vertices of G: for any vertex $a \in V(G)$, we define

$$R_a(v, x, y) = R(v, x, y),$$

with the stipulation that a is the root vertex. We similarly denote the corresponding combinatorial class by $\mathcal{R}$.

We denote by $M_a(x)$ the ordinary generating function of the number of paths in G which begin at vertex a. Since G is M-regular, we have

$$M_a(x) = \frac{1}{1 - xM}.$$

We then define the vector $\mathbf{M}(x, y)$, indexed by vertices $a \in V(G)$, by

$$\mathbf{M}_a(x, y) = M_a(x)M_a(y) = \frac{1}{(1 - xM)(1 - yM)},$$

so that

$$\mathbf{M}(x, y) = \frac{1}{(1 - xM)(1 - yM)}\mathbf{1},$$

where $\mathbf{1}$ is the vector of all 1s.

For $a \in V(G)$, $F_a^{k,k'}$ is the set of pairs of paths, the first having k distinct nodes and the other having k', which begin at a and never coincide thereafter (we call such a pair a *fork* and denote the combinatorial class by $\mathcal{F}$). Then $F(x, y)$ is given by

$$F_a(x, y) = \sum_{k,k' \geq 0} F_a^{k,k'} x^k y^{k'},$$

with an associated vector $\mathbf{F}(x, y)$, again indexed by vertices $a \in V(G)$.

For $(a, b) \in V(G)^2$, we denote by $\mathcal{N}_{a,b}^{k,k'}$ the set of pairs of paths with $k + 2$ and $k' + 2$ distinct vertices which both begin at a and end at b and which have no other common vertices. Then $N_{a,b}^{k,k'} = |\mathcal{N}_{a,b}^{k,k'}|$, and we then define the generating function at (a, b):

$$N_{a,b}(x, y) = \sum_{k,k' \geq 0} N_{a,b}^{k,k'} x^k y^{k'},$$

which gives an infinite-dimensional matrix $\mathbf{N}(x, y)$ whose entries are indexed by pairs of vertices in G. The associated combinatorial class we denote by $\mathcal{N}$. We refer to such pairs of paths as *bubbles*.

In the case where G is a translation-invariant graph, we can replace all matrices and vectors by simple functions in a natural way. In the special case of a grid graph (which, we note, is acyclic), since all paths between the same two points have the same length/number of distinct nodes, we can simplify our generating functions further: let $N(z)$ be the ordinary generating function of the number of pairs of paths starting at the root node and ending at a common terminal point with no intermediate points of contact; that is, N_k is the number of pairs of such paths with length $k + 1$. Furthermore, $B(z)$ is the generating function of B_k, the number of pairs of paths of length k starting at the root and ending at a common point. For general translation-invariant graphs, both of these generating functions must be replaced by bivariate ones, and we have the identity $N(x, y) = N(xy)$ in the grid case.

3.2 Proof Sketch We denote by $Q(z)$ the Poisson transform of the second moment of S_n:

$$Q(z) = \sum_{n \geq 0} \mathbb{E}[S_n^2] \frac{z^n}{n!} e^{-z},$$

which can be expressed as a summation over all potential trie nodes (i.e., compatible path/label pairs): denoting by $\mathcal{P}_k$ the set of paths of length k in G which begin at r,

$$Q(z) = \sum_{k,k' \geq 1} \sum_{P \in \mathcal{P}_k} \sum_{P' \in \mathcal{P}_{k'}} \sum_{w \in \mathcal{A}^k} \sum_{w' \in \mathcal{A}^{k'}} \Pr[(P, w), (P', w') \in GT(z)].$$

Here, the notation

$$\Pr[(P_1, w_1), \ldots, (P_j, w_j) \in GT(z)]$$

is shorthand for

$$\sum_{n \geq 0} \Pr\left[\forall_{i=1}^{j} (P_i, w_i) \in GT_n\right] \frac{z^n}{n!} e^{-z}.$$

We first derive an explicit expression for $Q(z)$ in terms of the sequences defined above. We claim that

$$Q(z) = \sum_{l,k,k' \geq 0} R_{l,k,k'} \sum_{w \in \mathcal{A}^k} \sum_{w' \in \mathcal{A}^{k'}} \left(\sum_{u \in \mathcal{A}^l} (\alpha\,(P(w)z) \star \alpha\,(P(w')z))(P(u)z) \right) + \sum_{u,u' \in \mathcal{A}^l, u \neq u'} \alpha\,(P(u)P(w)z)\,\alpha\,(P(u')P(w')z) \Bigg),$$

where $\star$ denotes the Poisson Hadamard product of two functions [4, 5], and $\alpha(z) = 1 - e^{-z} - ze^{-z}$. This expression follows easily from partitioning the set of pairs of rooted paths beginning at r according to the number of shared nodes; the partition elements are the $\mathcal{R}_{l,k,k'}$.

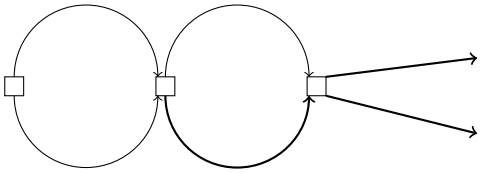

Figure 2: A schematic illustration of a pair of paths in $\mathcal{R}_{2,k,k'}$ and its decomposition into two bubbles and a fork.

In what follows, we develop the Hadamard product to make a tractable expression. We consider two paths P and P' which meet at exactly ℓ nodes. Consider also the strings u with $|u| = \ell$ (the labels for the common nodes of P and P') and w, w', which label the non-overlapping parts of P and P', respectively. Our goal is to determine an expression for

$$J(z) = \Pr[(P, uw), (P', uw') \in GT(z)]; \tag{3.1}$$

that is, the Poisson transform of the probability that, simultaneously,

$$(P, uw) \in GT_n$$

and

$$(P', uw') \in GT_n,$$

with u being the labels on the overlapping nodes. For this, we use the Poisson splitting property: there are N (a Poisson random variable with mean z) label functions, which may go into three buckets (with a fourth which is unimportant to the analysis):

- label functions which have u on the overlap points and are labelled by w on the remaining points of P but are not labelled with w' on the remainder of P'. This happens with probability $P(u)P(w)(1 - P(w'))$.
- label functions which have u on the overlap points and are labelled by w and w' on the remaining nodes of P and P', respectively. This happens with probability $P(u)P(w)P(w')$.
- label functions which have u on the overlap points and are labelled by w' on the remaining points of P' but are not labelled by w on the remainder of P. This happens with probability $P(u)(1 - P(w))P(w')$.

Now, in order to determine an expression for (3.1), we condition on the number of label functions that fall into the middle bucket.

$$\begin{aligned}
J(z) = {} & (1 - e^{-P(uww')z} - P(uww')ze^{-P(uww')z}) \\
& + P(uww')ze^{-P(uww')z} \\
& \cdot (1 - e^{-P(uw)(1-P(w'))z})(1 - e^{-P(u)(1-P(w))P(w')z}) \\
& + e^{-P(uww')z} \cdot (1 - e^{-P(uw)(1-P(w'))z} \\
& - P(uw)(1 - P(w'))ze^{-P(uw)(1-P(w'))z}) \\
& (1 - e^{-P(uw')(1-P(w))z} \\
& - P(uw')(1 - P(w))ze^{-P(uw')(1-P(w))z}).
\end{aligned}$$

To form the Poisson variance

$$V(z) = Q(z) - (S(z))^2,$$

we add the term

$$\begin{aligned}
& -(1 - e^{-P(uw)z} - P(uw)ze^{-P(uw)z}) \\
& \quad \cdot (1 - e^{-P(uw')z} - P(uw')ze^{-P(uw')z})
\end{aligned}$$

to the above expression. Simplifying, we find terms containing factors of the form

$$e^{-P(uw)z - P(uw')z + P(uww')z},$$

which cannot be handled in an obvious way by Mellin transforms. We thus define $\beta = P(ww')$, which yields

$$\begin{aligned}
J(z) = {} & e^{-\beta P(u)z} \\
& \left(1 - ((P(w) - \beta)P(u)z + 1)e^{-(P(w)-\beta)P(u)z}\right) \\
& \cdot \left(1 - ((P(w') - \beta)P(u)z + 1)e^{-(P(w')-\beta)P(u)z}\right) \\
& + \beta P(u)ze^{-\beta P(u)z}\left(1 - e^{-(P(w)-\beta)P(u)z}\right) \\
& \cdot \left(1 - e^{-(P(w')-\beta)P(u)z}\right) \\
& + 1 - (\beta P(u)z + 1)e^{-\beta P(u)z} \\
& - \left(1 - (P(wu)z + 1)e^{-P(wu)z}\right) \\
& \cdot \left(1 - (P(w'u)z + 1)e^{-P(w'u)z}\right)
\end{aligned}$$

Then, Taylor expanding with respect to $\beta \to 0$, we find

$$\begin{aligned}
J(z) = {} & z^3 P(u^3 ww')\beta e^{-(P(w)+P(w'))P(u)z} \\
& + z^2 P(u^2)\beta^2(1 - P(uw)z) \\
& \cdot (1 - P(uw')z)e^{-(P(w)+P(w'))P(u)z} \\
& + O(z^3\beta^3 e^{-(P(w)+P(w')-\beta)P(u)z}).
\end{aligned}$$

The error term is still difficult to handle exactly, so we upper bound it:

$$\begin{aligned} P(w) + P(w') - \beta &= (P(w) - \beta) + (P(w') - \beta) + \beta \\ &\geq \frac{P(w) + P(w')}{2} - \beta + \beta \\ &= \frac{P(w) + P(w')}{2}, \end{aligned}$$

which yields

$$O(z^3\beta^3 e^{-\frac{(P(w)+P(w'))P(u)z}{2}}). \tag{3.2}$$

Each term is now amenable to analysis via double Mellin transforms [7].

For the first term $J_1(z)$, we define

$$J_1(z_1, z_2) = P(u^2w)z_1 e^{-P(uw)z_1} P(uw')z_2 e^{-P(uw')z_2}$$

and note that

$$J_1(z) = zJ_1(z, z).$$

Then we take the Mellin transform with respect to both z_1 and z_2:

$$J_1^*(s_1, s_2) = \int_{\mathcal{C}_1}\int_{\mathcal{C}_2} J_1(z_1, z_2) z_2^{s_2-1} z_1^{s_1-1}\, \mathrm{d}z_2\, \mathrm{d}z_1,$$

with $\mathcal{C}_1, \mathcal{C}_2$ arbitrary contours starting at the origin and tending to ∞ in a cone around the positive real axis. This yields

$$J_1^*(s_1, s_2) = (P(uw))^{-s_1-s_2+1}\Gamma(s_1 + 1)\Gamma(s_2 + 1).$$

This results (after multiplying by $R_{l,k,k'}$ and summing) in the expression

$$V_1(s_1, s_2) = R_r(H(s_1 + s_2 - 1), H(s_1 - 1), H(s_2 - 1)) \tag{3.3}$$
$$\cdot \Gamma(s_1 + 1)\Gamma(s_2 + 1), \tag{3.4}$$

where we recall that $R_r(v, x, y)$ denotes the component of the vector $\mathbf{R}(v, x, y)$ in which the root vertex is r.

To characterize the singularities of this function, we shall need some combinatorial identities relating $\mathbf{R}$ to $\mathbf{N}$ and $\mathbf{F}$. For definitions of these objects, see Section 3.1.

LEMMA 3.1. (IDENTITIES FOR $\mathbf{R}$, $\mathbf{N}$, AND $\mathbf{F}$) *We have*

$$\mathbf{R}(v, x, y) = (\mathbf{I} - v\mathbf{N}(x, y))^{-1}\mathbf{F}(x, y), \tag{3.5}$$

$$(\mathbf{I} - xy\mathbf{N}(x, y))^{-1}\mathbf{F}(x, y) = \mathbf{M}(x, y) \tag{3.6}$$
$$= \frac{1}{(1 - xM)(1 - yM)}\mathbf{1}, \tag{3.7}$$

where $\mathbf{I}$ *denotes the identity matrix.*

Proof. $\mathbf{R}$ IDENTITY
To prove this, we note that an element of R is simply a finite sequence of elements of $\mathcal{N}$, followed by an element of $\mathcal{F}$. Each pair of paths in $\mathcal{N}$ contributes one new point of coincidence (the shared terminal node of the two paths), so we must multiply by v. All of this gives an equation on generating functions, which is precisely the claimed identity.

$\mathbf{M}$ IDENTITY
The proof of the second identity uses a decomposition that is quite similar to the one in the first: a rooted pair of paths, represented by $\mathbf{M}(x, y)$, is a sequence of bubbles followed by a fork. Now, the coefficient $[x^k y^{k'}]N_{a,b}(x, y)$ is, by definition, the number of pairs of paths from a to b in the class $\mathcal{N}$, which have *length* (i.e., number of edges) $(k+1, k'+1)$. Such a pair of paths then has a *size* (i.e., number of vertices) $(k + 2, k' + 2)$. We therefore mark the beginning of each pair of paths with multiplication by xy. Putting everything together yields the claimed equation.

Applying this lemma and setting $v = H(s_1+s_2-1)$, $x = H(s_1 - 1)$, and $y = H(s_2 - 1)$, we can rewrite (3.3) as

$$\begin{aligned} V_1^*(s_1, s_2) =&(\Gamma(s_1 + 1)\Gamma(s_2 + 1)\cdot \\ &(\mathbf{I} - v\mathbf{N}(x, y))^{-1}(\mathbf{I} - xy\mathbf{N}(x, y))\mathbf{M}(x, y))_r. \end{aligned}$$

Now we invoke the hypothesis that the distribution over the alphabet is uniform. This implies a simple expression for $H(s)$:

$$H(s) = A^{s+1}.$$

We then see that $v = H(s_1 + s_2 - 1) = A^{s_1+s_2} = H(s_1 - 1)H(s_2 - 1) = xy$. Thus,

$$(\mathbf{I} - v\mathbf{N}(x, y))^{-1}(\mathbf{I} - xy\mathbf{N}(x, y)) = \mathbf{I}.$$

This leaves

$$V_1^*(s_1, s_2) = \frac{\Gamma(s_1 + 1)\Gamma(s_2 + 1)}{(1 - MH(s_1 - 1))(1 - MH(s_2 - 1))}.$$

This formula depends only on M and no other structural property of G or on the choice of root node.

This function is easily seen to be meromorphic in both variables, with infinitely many simple poles at

$$s_1 = -\rho + 1 + \frac{2\pi i j}{\log|\mathcal{A}|} \qquad j \in \mathbb{Z}$$
$$s_2 = -\rho + 1 + \frac{2\pi i k}{\log|\mathcal{A}|} \qquad k \in \mathbb{Z}.$$

We are now ready to recover asymptotics of $V_1(z,z)$ via the *double* inverse Mellin transform:

$$V_1(z_1,z_2) = \frac{1}{(2\pi i)^2}\int_{c_1-i\infty}^{c_1+i\infty} \int_{c_2-i\infty}^{c_2+i\infty} V^*(s_1,s_2) z^{-s_1-s_2}\,\mathrm{d}s_1\,\mathrm{d}s_2,$$

where the contours of integration are chosen to lie in the fundamental strip of $V_1^*(s_1,s_2)$. Both integrals can be handled via their singularities; that is, we apply Theorem 4 of [3], which yields the estimate

$$V_1(z) = O(z^{2\rho-1}),$$

though we will see that, after de-Poissonization, cancellation of terms yields only a contribution of $\Theta(n^\rho)$.

We analyze the second term, $J_2(z)$, along similar lines, defining

$$\begin{aligned} J_2(z_1,z_2) =& z_1 z_2 P(u^2)\beta^2(1-P(uw)z_1) \\ & \cdot (1-P(uw')z_2)e^{-P(uw)z_1-P(uw)z_2} \end{aligned}$$

and noting that $J_2(z) = J_2(z,z)$. After taking the double Mellin transform, we find, with $V_2(s_1,s_2)$ defined analogously to $V_1(s_1,s_2)$ above,

$$V_2(s_1,s_2) = R_r(A^{1+s_1+s_2}, A^{s_1}, A^{s_2}) s_1^2 s_2^2 \Gamma(s_1)\Gamma(s_2).$$

Note that, in this case, $v \neq xy$, so that the convenient cancellation that occurred with the first term does not happen here. Applying Lemma 3.1, this becomes

$$\begin{aligned} V_2(s_1,s_2) =& ((\mathbf{I}-v\mathbf{N}(x,y))^{-1}(\mathbf{I}-xy\mathbf{N}(x,y))\mathbf{M}(x,y))_r \\ & s_1^2 s_2^2 \Gamma(s_1)\Gamma(s_2), \end{aligned}$$

which is singular when $s_1+s_2 = -\rho'$, for some ρ' which is the supremum of all values for which

$$\sum_{n=0}^{\infty} (A^{1-\rho'}\mathbf{N}(A^{s_1},A^{s_2}))^n = \infty. \tag{3.8}$$

We will show that $\rho' \geq \rho$: we can do this by showing that $A^{1-\rho'} \leq \frac{1}{M}$, for then

$$\rho' \geq 1 + \frac{\log M}{\log A} = \rho.$$

To do this for arbitrary directed, acyclic M-regular graphs, let $\rho_*(\mathbf{N}(A^{s_1},A^{s_2}))$ denote the spectral radius of $\mathbf{N}(A^{s_1},A^{s_2})$. Since the entries of $\mathbf{N}(x,y)$ are greater than or equal to those of the adjacency matrix of G whenever $x,y>0$, we have

$$\rho_*(\mathbf{N}(A^{s_1},A^{s_2})) \geq \rho_*(G),$$

where $\rho_*(G)$ is the spectral radius of the adjacency matrix of G, and since G is M-regular, $\rho_*(G) = M$. Now, let s_1, s_2 be such that

$$A^{1+s_1+s_2}M = 1,$$

so that, in particular, the equality (3.8) holds. Then

$$-(s_1+s_2) \leq \rho',$$

since ρ' is the supremum of values for which the sum diverges, which yields

$$A^{1-\rho'}M \leq 1,$$

as desired.

This proves the claimed inequality, but, since we have not ruled out the possibility of non-simple poles, this shows only that this term makes a contribution of $\Omega(n^{\rho'})$. For translation-invariant and time-expanded graphs (which are of practical interest), we can show more: roughly speaking, the invariance under certain translations exhibited by graphs in both classes allows us to replace the infinite matrices used above by finite ones, and then we can apply the Perron-Frobenius theorem to conclude that there is a spectral gap, which implies that the pole which determines the asymptotics is simple. In this case, this implies that the contribution of the second term is $\Theta(n^{\rho'})$.

We can calculate ρ' explicitly for the special case of a square grid, with $M=2$. In this case, since G is translation invariant, it is easier to resort to simple generating functions (rather than matrices and vectors). Recall the definitions of $N(z)$ and $B(z)$. For general grids, we have the identity

$$N(z) = \frac{1}{z}\left(1 - \frac{1}{B(z)}\right),$$

whose proof is very similar to that of Lemma 3.1. Now, since $M=2$, we have a nice formula for $B(z)$ (which we do not prove, for lack of space): $B(z) = \frac{1}{\sqrt{1-4z}}$. Plugging in $z = A^{-\rho'}$ and solving for ρ' gives the claimed value.

The error term (3.2) can be handled in a manner similar to the handling of the second term. We omit the details here, but it turns out that it is of order $\Theta(n^{\rho''})$, where $\rho'' < \rho'$.

It remains to transfer asymptotics from $V(z)$ to V_n. Roughly speaking, the plan is to apply the de-Poissonization theorems of [6] to $Q(z)$ (recalling that this is the Poisson generating function of the second moment of S_n, then that $V_n = \mathbb{E}[S_n^2] - \mathbb{E}[S_n]^2$). We omit the details for lack of space, but we note that, as in the case where G is a tree, the first term of order $\Theta(z^{2\rho-1})$

in the resulting expansion cancels in the uniform case, yielding a contribution of order $O(n^{\rho})$. The $\Theta(n^{\rho'})$ contribution of the second term is not cancelled, so we end up with

$$V_n = \Theta(n^{\rho'}).$$

4 Extension to Non-Uniform Alphabet

The case in which the distribution over the alphabet is non-uniform is more interesting, both in the phenomena exhibited and the challenges in the analysis. Since, in general, $H(s_1 - 1)H(s_2 - 1) \neq H(s_1 + s_2 - 1)$, the cancellation exhibited in the uniform case does not hold. Thus, the properties of $\mathbf{N}(x, y)$ become important, and the order of the variance changes with G (it is minimal when G is a tree).

As a simple example, consider the case where G is M-regular and translation invariant. In this case, the Mellin transform formula becomes

(4.9)
$$V^*(s_1, s_2) = \frac{\Gamma(s_1 + 1)\Gamma(s_2 + 1)}{B(xy)(1 - xM)(1 - yM)(1 - vN(xy))},$$

where $v = H(s_1 + s_2 - 1)$, $x = H(s_1 - 1)$, and $y = H(s_2 - 1)$. Using the identity on N and B mentioned earlier, the singularity analysis of (4.9) boils down to the study of $B(z)$. We further restrict our attention to the case where $M = 2$ or 3, since then $B(M^{-2}) = \infty$. Now, for $s_1 = s_2$, we have

$$H(s_1 - 1)H(s_2 - 1) < H(s_1 + s_2 - 1),$$

by convexity of the power function. Thus, we find that there exists some $-\rho_2 < -\rho+1$ for which $1-vN(xy) = 0$ whenever $x = y = -\rho_2$, and this is the furthest singularity of $V^*(s_1, s_2)$ to the left and is simple. Thus, with some additional complications, we find that the polynomial order of growth of V_n is $2\rho_2 + 1$, which is larger than the $2\rho - 1$ in the tree case. For $M > 3$, $B(M^{-2}) < \infty$, and the singularity analysis is more intricate.

References

[1] Jérémie Bourdon. Size and path length of patricia tries: Dynamical sources context. *Random Struct. Alg.*, 19:289–315, 2001.

[2] Michael Drmota and Wojciech Szpankowski. The expected profile of digital search trees. *J. Comb. Theory Ser. A*, 118(7):1939–1965, October 2011.

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

[4] H.K. Hwang, M. Fuchs, and V. Zacharovas. Asymptotic variance of random symmetric digital search trees. *Discrete Mathematics & Theoretical Computer Science*, 12(2):103–166, 2010.

[5] Philippe Jacquet. Trie structure for graph sequences. *25th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms*, 2014.

[6] Philippe Jacquet and Wojciech Szpankowski. Analytical depoissonization and its applications. *Theor. Comput. Sci.*, 201(1-2):1–62, July 1998.

[7] Philippe Jacquet and Wojciech Szpankowski. Joint string complexity for Markov sources. *DMTCS Proceedings*, 01:303–322, 2012.

[8] M. Regnier and P. Jacquet. New results on the size of tries. *Information Theory, IEEE Transactions on*, 35(1):203–205, 1989.