A New Binomial Recurrence Arising in a Graphical Compression Algorithm

Yongwook Choi[1], Charles Knessl[2†] and Wojciech Szpankowski[3‡]

[1] J. Craig Venter Institute, USA
[2] Department of Mathematics, Statistics & Computer Science, University of Illinois at Chicago, USA
[3] Department of Computer Science, Purdue University, USA

In a recently proposed graphical compression algorithm by Choi and Szpankowski (2009), the following tree arose in the course of the analysis. The root contains n balls that are consequently distributed between two subtrees according to a simple rule: In each step, all balls independently move down to the left subtree (say with probability p) or the right subtree (with probability $1-p$). A new node is created as long as there is at least one ball in that node. Furthermore, a nonnegative integer d is given, and at level d or greater one ball is removed from the leftmost node before the balls move down to the next level. These steps are repeated until all balls are removed (i.e., after $n+d$ steps). Observe that when $d=\infty$ the above tree can be modeled as a *trie* that stores n independent sequences generated by a memoryless source with parameter p. Therefore, we coin the name (n,d)-tries for the tree just described, and to which we often refer simply as d-tries. Parameters of such a tree (e.g., path length, depth, size) are described by an interesting two-dimensional recurrence (in terms of n and d) that – to the best of our knowledge – was not analyzed before. We study it, and show how much parameters of such a (n,d)-trie differ from the corresponding parameters of regular tries. We use methods of analytic algorithmics, from Mellin transforms to analytic poissonization.

Keywords: Digital trees, Mellin transform, poissonization, graph compression

1 Introduction

In [1] an algorithm was described to compress the *structure* of a (unlabeled) graph. The main idea behind the algorithm is quite simple: First, a vertex of a graph, say v_1, is selected and the *number* of neighbors of v_1 is stored in a binary string. Then the remaining $n-1$ vertices are partitioned into two sets: the neighbors of v_1 and the non-neighbors of v_1. This process continues by selecting randomly a vertex, say v_2, from the neighbors of v_1 and storing two *numbers*: the number of neighbors of v_2 among each of the above two sets. Then the remaining $n-2$ vertices are partitioned into four sets: the neighbors of both v_1 and v_2, the neighbors of v_1 that are non-neighbors of v_2, the non-neighbors of v_1 that are neighbors of

† This author was supported by NSA Grant H98230-11-1-0184.
‡ This work was supported in part by the NSF Science and Technology Center for Science of Information Grant CCF-0939370, NSF Grants CCF-0830140 and DMS-0800568, AFOSR Grant FA8655-11-1-3076, NSA Grant H98230-11-1-0141, and the MNSW Grant N206 369739.

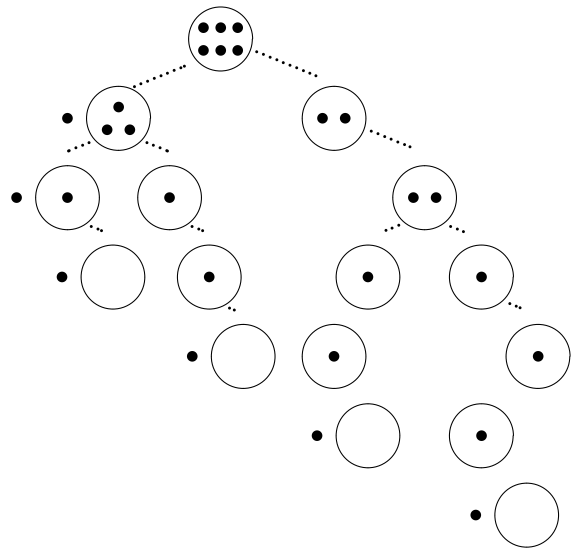

Fig. 1: A $(6, 1)$-trie with six balls and $d = 1$, in which the deleted ball is shown next to the node where it was removed.

v_2, and the non-neighbors of both v_1 and v_2. This procedure continues until all vertices are processed. In the Erdős-Rényi model, a random graph has any pair of vertices connected by an edge with probability p. It is proved in [1] that for large n our algorithm optimally compresses any (unlabeled) graph generated by the Erdős-Rényi model (and, in fact, it works well in practice even for graphs not generated by the Erdős-Rényi model). To establish this asymptotic optimality result, an interesting tree was used in the construction described next.

The root of such a tree contains n balls (vertices of the underlying graph) that are consequently distributed between two subtrees according to a simple rule: In each step, all balls independently move down to the left subtree (say with probability p) or the right subtree (with probability $1 - p$), and a new node is created as long as there is at least one ball in that node. Finally, a non-negative integer d is given so that at level d or greater one ball is removed from the leftmost node before the balls move down to the next level. These steps are repeated until all balls are removed (i.e., after n+d steps). Of interest are such tree parameters as the depth, path length (sum of all depths), size, and so forth. This is illustrated in Figure 1 in which the deleted ball is shown next to the node from where it was removed.

The tree just described falls between two digital trees, namely *tries* and *digital search trees* [3, 12, 14, 19]. In fact, when $d = \infty$ the tree can be modeled as a *trie* that stores n independent sequences generated by a memoryless source with parameter p. Hence, we coin the term (n, d)-trie (or simply d-trie) for the tree just described. In [1] lower and upper bounds were proved for parameters of interest, by using known results for tries and digital search trees [3, 19]. In this paper, we establish precise asymptotic results. In particular, we show by how much the path length of a d-trie differs from the path length of the corresponding regular trie.

Many parameters of a (n, d)-trie can be described by the following two dimensional recurrence

$$a(n,d) = f(n) + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} [a(k,d-1) + a(n-k,k+d-1)], \quad d \geq 1, \tag{1}$$

and the boundary equation

$$a(n+1,0) = f(n) + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} [a(k,0) + a(n-k,k)], \tag{2}$$

for $0 < p < 1$, $q = 1-p$, and a known additive term $f(n)$. For example, when $f(n) = n$, then $a(n,d)$ represents the path length. Recurrence (2) is equivalent to the following boundary condition

$$a(n,1) = a(n+1,0).$$

For $d = \infty$ recurrence (1) becomes a traditional recurrence arising in the analysis of tries [19] whose solutions (exact and asymptotic) are well known. Thus, it is natural to study the difference $\tilde{a}(n,d) := a(n,d) - a(n,\infty)$, and that is our objective. In passing, we should point out that recurrence (2) resembles the one used to analyze another digital search tree, known as a *digital search tree*. In this paper we prove, however, that a (n,d)-trie more closely resembles a trie, rather than a digital search tree.

Our main interest lies in solving recurrence (1) for fixed d. For graph compression we only need $d = 0$, and we focus on this case. In particular, for $f(n) = n$ (that is, for the path length in a d-trie) we prove that the excess quantity $\tilde{a}(n,d)$ becomes asymptotically, as $n \to \infty$ and $d = O(1)$,

$$\frac{1}{2h\log p} \log^2 n + \frac{d}{h} \log n + \left[-\frac{1}{2h} + \frac{1}{h \log p} \left(\gamma + 1 + \frac{h_2}{2h} + \Psi(\log_p n) \right) \right] \log n$$

where $\Psi(\cdot)$ is the periodic function when $\log p / \log(1-p)$ is rational, and h is the entropy rate.

Digital trees such as tries and digital search trees have been intensively studied for the last thirty years [2, 3, 5, 7, 11, 12, 13, 16, 17, 18, 19]. However, our two-dimensional recurrence seems to be new and harder to analyze. It somewhat resembles the profile recurrences for digital trees, which were studied for tries in [15] and digital search trees in [4], and which are known to also be challenging.

The paper is organized as follows. In the Section 2 we precisely formulate our problem and analyze it for $f(n) = n$. Some proofs are presented in Section 3, while further details are provided in our journal version of this paper.

2 Problem Statement

In this section, we first formulate some recurrences describing (n,d)-tries, then summarize our main results and discuss some extensions.

2.1 Main Results

Let us consider a (n,d)-trie with n balls and parameter $d \geq 0$. First, we analyze the average path length $b(n,d)$. It satisfies the following recurrence equations

$$b(n+1,0) = n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left[b(k,0) + b(n-k,k) \right], \quad \text{for } n \geq 2, \tag{3}$$

and

$$b(n,d) = n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left[b(k, d-1) + b(n-k, k+d-1) \right], \quad \text{for } n \geq 2, d \geq 1. \tag{4}$$

Recurrence (3) follows from the fact that starting with $n+1$ balls in the root node, and removing one ball, we are left with n balls passing through the root node. The root contributes n since each time a ball moves down it adds 1 to the path length. Those n balls move down to the left or the right subtrees. Let us assume k balls move down to the left subtree (the other $n-k$ balls must move down to the right subtree); this occurs with probability $\binom{n}{k} p^k q^{n-k}$. At level one, one ball is removed from those k balls in the root of the left subtree. This contributes $b(k,0)$. There will be no removal from $n-k$ balls in the right subtree until all k balls in the left subtree are removed. This contributes $b(n-k,k)$. Similarly, for $d > 0$ we arrive at recurrence (4).

Here $0 < p < 1$ and $q = 1-p$, and we also use the boundary conditions

$$b(0,d) = b(1,d) = 0, \quad d \geq 0; \quad b(2,0) = 0. \tag{5}$$

By setting $d = 1$ in (4) and comparing the result to (3) we can replace (3) by the simpler boundary condition

$$b(n,1) = b(n+1,0), \quad \text{for } n \geq 0. \tag{6}$$

We are primarily interested in estimating $b(n,0)$ for large n.

If we let $d \to \infty$ in (4) and assume that $b(n,d)$ tends to a limit $b(n,\infty)$, then (4) becomes

$$b(n,\infty) = n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left[b(k,\infty) + b(n-k,\infty) \right], \tag{7}$$

with $b(0,\infty) = b(1,\infty) = 0$. This is the same as the recurrence for the mean path length in a trie, which was analyzed, for example, in [12, 19]. One form of the solution is given by the alternating sum

$$b(n,\infty) = \sum_{\ell=2}^{n} (-1)^\ell \binom{n}{\ell} \frac{\ell}{1-p^\ell-q^\ell}, \tag{8}$$

and an alternate form is given by the integral [19]

$$b(n,\infty) = \frac{n!}{(2\pi i)^2} \oint \left[\int_{Br} z^{-s} \frac{\Gamma(s+1)}{1-p^{-s}-q^{-s}} ds \right] \frac{e^z}{z^{n+1}} dz, \tag{9}$$

where $\Gamma(\cdot)$ is the Gamma function, Br is a vertical Bromwich contour on which $-2 < \Re(s) < -1$ and the z-integral is over a small loop about $z = 0$.

The asymptotic expansion of (9) as $n \to \infty$ may be obtained by a combination of singularity analysis and depoissonization arguments (see [7, 9, 19]) and we obtain

$$b(n,\infty) = \frac{1}{h} n \log n + \frac{1}{h} \left[\gamma + \frac{h_2}{2h} + \Phi(\log_p n) \right] n + o(n), \tag{10}$$

where $h = -p\log p - q\log q$, $h_2 = p\log^2 p + q\log^2 q$, γ is the Euler constant, and $\Phi(x)$ is the periodic function

$$\Phi(x) = \sum_{k=-\infty, k\neq 0}^{\infty} \Gamma\left(-\frac{2k\pi i r}{\log p}\right) e^{2k\pi r i x}, \tag{11}$$

provided that $\log p/\log q = r/s$ is rational, with r and s being integers with $\gcd(r,s) = 1$. If $\log p/\log q$ is irrational, then the term with Φ is absent from the $O(n)$ term of (10). We shall later use (10) to analyze the behavior of $b(n,d)$ for $n \to \infty$ and a fixed d.

Next we set

$$\tilde{b}(n,d) = b(n,d) - b(n,\infty) \tag{12}$$

so that $\tilde{b}(n,d)$ measures how the path lengths in the d-trie differs from those in a trie. From (4) and (7), we then obtain

$$\tilde{b}(n,d) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left[\tilde{b}(k,d-1) + \tilde{b}(n-k,k+d-1)\right], \quad \text{for } n \geq 2,\, d \geq 1, \tag{13}$$

which unlike (4) is a homogeneous recurrence. Then from (6) and (12) we have the boundary condition

$$\tilde{b}(n+1,0) - \tilde{b}(n,1) = b(n,\infty) - b(n+1,\infty). \tag{14}$$

From (5) and (7) we also have $\tilde{b}(0,d) = \tilde{b}(1,d) = 0$ for $d \geq 0$.

We further define $b_*(n,d)$ to be the solution of

$$b_*(n,d) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} b_*(k,d-1), \quad \text{for } n \geq 2,\, d \geq 1, \tag{15}$$

and

$$b_*(n+1,0) - b_*(n,1) = b(n,\infty) - b(n+1,\infty). \tag{16}$$

Note that (15) differs from (13) in that the former neglects the term involving $\tilde{b}(n-k,k+d-1)$. We will show that this term in (13) is asymptotically negligible for $n \to \infty$ with fixed d, so that $\tilde{b}(n,d) \sim b_*(n,d)$. The recurrence (15) is much easier to solve by transform methods [7, 19] than is (13).

We summarize our main result below. In Section 3 we establish Theorem 1 along with some other exact and asymptotic results for (3)-(6) and (13)-(16).

Theorem 1 *For $n \to \infty$ and $d = O(1)$ we have $\tilde{b}(n,d) = O(\log^2 n)$. More precisely*

$$\tilde{b}(n,d) = \frac{1}{2h\log p}\log^2 n + \frac{d}{h}\log n + \left[-\frac{1}{2h} + \frac{1}{h\log p}\left(\gamma + 1 + \frac{h_2}{2h} + \Psi(\log_p n)\right)\right]\log n + O(1), \tag{17}$$

where $\Psi(\cdot)$ is the periodic function

$$\Psi(x) = \sum_{k=-\infty, k\neq 0}^{\infty} \left[1 + \frac{2k\pi i r}{\log p}\right] \Gamma\left(-\frac{2k\pi i r}{\log p}\right) e^{2k\pi i r x} \tag{18}$$

and $\log p/\log q = r/t$ is rational, as in (11). If $\log p/\log q$ is irrational, the term involving Ψ in (17) is absent.

We see that $b(n,d) - b(n,\infty) = O(\log^2 n)$, which shows that the (n,d)-tries studied in [1] are in some sense more similar to tries than to digital search trees (DST). In [1], it was shown that $b(n,0)$ was bounded above by average path lengths in tries and below by average path lengths in DST's. It was also conjectured that $b(n,d) - b(n,\infty)$ is $O(n)$, but our result shows that this difference is in fact much smaller.

2.2 Related Recurrence Equations

The method presented in the next section allow us to analyze a class of recurrences of the type (3) with inhomogeneous terms other than n. For example, suppose we define $a(n,d)$ by

$$a(n,d) = f(n) + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} [a(k,d-1) + a(n-k,k+d-1)] \tag{19}$$

where $f(n)$ is a given sequence that grows algebraically or logarithmically for $n \to \infty$. The boundary condition is again of the type (3), or equivalently,

$$a(n,1) = a(n+1,0), \tag{20}$$

and we have $a(0,d) = a(1,d) = 0$. Also, let $a(n,\infty)$ satisfy (19) with the second argument of $a(\cdot,\cdot)$ replaced by infinity. This recurrence can be solved by generating functions and Mellin transforms, and we can then establish that $a(n,d) - a(n,\infty) \equiv \tilde{a}(n,d)$, will satisfy

$$\tilde{a}(n,d) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} [\tilde{a}(k,d-1) + \tilde{a}(n-k,k+d-1)] \tag{21}$$

and

$$\tilde{a}(n+1,0) - \tilde{a}(n,1) = a(n,\infty) - a(n+1,\infty). \tag{22}$$

The asymptotic behavior of $\tilde{a}(n,d)$ for $d = O(1)$ and $n \to \infty$ can be obtained in a manner completely analogous to the case $f(n) = n$, discussed in the next section.

For example, the case

$$f(n) = \lceil \log(n+1) \rceil$$

arose in analyzing the compression algorithm in [1]. In [1] it was shown that $a(n,\infty)$ has the asymptotic form

$$a(n,\infty) = \frac{n}{h} A_*(-1) + o(n), \quad n \to \infty \tag{23}$$

where

$$A_*(s) = \sum_{k \geq 2} \frac{\lceil \log(k+1) \rceil}{k!} \Gamma(k+s).$$

if $\log p / \log q$ is irrational. If $\log p / \log q = r/s$ is rational, the constant $A_*(-1)$ in (23) must be replaced by the oscillatory function

$$A_*(-1) + \sum_{k=-\infty, k\neq 0}^{\infty} A_* \left(-1 + \frac{2k\pi i r}{\log p} \right) e^{2k\pi i r \log_p n}. \tag{24}$$

By analyzing (21) and (22) for $n \to \infty$ we can show that the difference $a(n,d) - a(n,\infty)$ is $O(\log n)$, and more precisely

$$\tilde{a}(n,0) = a(n,0) - a(n,\infty) \sim \frac{A_*(-1)}{h \log p} \log n.$$

Again if $\log p / \log q$ is rational we must replace $A_*(-1)$ by the Fourier series in (24).

3 Analysis

We first give an intuitive derivation of the asymptotics of $b(n,d)$ for fixed $d \geq 0$ and $n \to \infty$, and in particular of $b(n,0)$.

We use the fact that for algebraically or logarithmically varying smooth $f(k)$ (for $k \to \infty$) we have (see [6, 10] for rigorous proofs)

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} f(k) = f(np) + O(nf''(np)),\ n \to \infty. \tag{25}$$

Starting from (13) we argue that the second sum is negligible for $n \to \infty$. Then if $\tilde{b}(n,d)$ varies weakly with n, we use (25) to approximate the first sum, which will be asymptotic to $\tilde{b}(np, d-1)$ so that (13) becomes

$$\tilde{b}(n,d) \sim \tilde{b}(np, d-1),\ n \to \infty \tag{26}$$

and, in particular,

$$\tilde{b}(n,1) \sim \tilde{b}(np,0),\ n \to \infty \tag{27}$$

which when added to (14) leads to

$$\tilde{b}(n+1,0) - \tilde{b}(np,0) \sim b(n,\infty) - b(n+1,\infty). \tag{28}$$

The right side of (28) may be estimated from (10) or by (9). Using (9) we can show that term by term differentiating of the asymptotic series in (10) is permissible, and thus (28) becomes, for $n \to \infty$,

$$\tilde{b}(n+1,0) - \tilde{b}(np,0) = -\frac{1}{h} \log n - \frac{1}{h}\left(\gamma + 1 + \frac{h_2}{2h}\right) - \frac{1}{h}\psi(\log_p n) + o(1), \tag{29}$$

where $\psi(\cdot)$ is the periodic function

$$\psi(x) = \sum_{k=-\infty, k\neq 0}^{\infty} \left[1 + \frac{2k\pi i r}{\log p}\right] \Gamma\left(-\frac{2k\pi i r}{\log p}\right) e^{2k\pi i r x}, \tag{30}$$

where we note that, in view of (11), $\psi(x) = \Phi(x) + (\log p)^{-1}\Phi'(x)$.

Now (29) suggests that $\tilde{b}(n,0)$ admits an asymptotic expansion of the form

$$\tilde{b}(n,0) = A \log^2 n + B \log n + C + o(1),\ n \to \infty \tag{31}$$

and then

$$\tilde{b}(n+1,0) - \tilde{b}(np,0) = -2A(\log p)\log n - A\log^2 p - B\log p + o(1). \tag{32}$$

Comparing (29) to (32) we conclude that $A = (2h \log p)^{-1}$ and then

$$B = -\frac{1}{2h} + \frac{1}{h \log p}\left[\gamma + 1 + \frac{h_2}{2h} + \psi(\log_p n)\right]. \tag{33}$$

We have thus formally derived the result in Theorem 1 for $\tilde{b}(n,0)$. For any fixed $d > 0$ we can extend this argument by asymptotically solving (26) by an expansion of the form

$$\tilde{b}(n,d) = A(d)\log^2 n + B(d)\log n + O(1) \tag{34}$$

to find from (26) that $A(d) = A(d-1)$ and $B(d) = B(d-1) + 2\log p A(d-1)$. Then using (34) in (28) or (29) we find that $A(d) = A(0) = (2h\log p)^{-1}$ and $B(d) - B(d-1) = 2\log p A(d-1) = h^{-1}$ so that $B(d) = B(0) + h^{-1}d$, where $B(0) = B$ is as in (33).

We proceed to provide a more rigorous derivation of the theorem. We first inductively establish the bound

$$\tilde{b}(n,d) \le A_0 n^{\nu+\epsilon}(p^2+q^2)^d;\ n \ge 2, d \ge 0 \tag{35}$$

where $\nu = \log(p^2+q^2)/\log(p)$, and this holds for all $\epsilon > 0$. When $n = 2$ we have (exactly) $\tilde{b}(2,d) = (2 - \frac{1}{pq})(p^2+q^2)^{d-1}$ so (35) clearly holds.

Assuming that (35) holds for all (N, D) with $N + D < n + d$, we can estimate the first sum in (13) by

$$\begin{aligned}\sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}\tilde{b}(k,d-1) &\le \sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}A_0 k^{\nu+\epsilon}(p^2+q^2)^{d-1}\\ &\le A_0(np)^{\nu+\epsilon}(p^2+q^2)^{d-1} = A_0 n^{\nu+\epsilon}p^{\epsilon}(p^2+q^2)^d.\end{aligned}$$

Using the inductive assumption in (35) we then estimate the second sum in (13) by

$$\begin{aligned}\sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}\tilde{b}(n-k,k+d-1) &\le A_0\sum_{k=0}^{n}(n-k)^{\nu+\epsilon}(p^2+q^2)^{k+d-1}p^k q^{n-k}\binom{n}{k}\\ &\le A_0 n^{\nu+\epsilon}(p^2+q^2)^{d-1}\sum_{k=0}^{n}\binom{n}{k}\left[p(p^2+q^2)\right]^k q^{n-k}\\ &= A_0 n^{\nu+\epsilon}(p^2+q^2)^{d-1}\left[q + p(p^2+q^2)\right]^n\end{aligned}$$

which is smaller than the first sum, by an exponentially small factor.

Combining these estimates and recalling that ϵ can be made arbitrarily small leads to the conclusion that (35) holds by induction.

By subtracting (15) from (13) and using the estimate in (35) to bound the second sum in the right side of (13), we conclude that $\tilde{b}(n,d) - b_*(n,d) = O(1)$ for $n \to \infty$. We proceed to analyze (15), with (16), and thus re-establish Theorem 1.

Introducing the exponential generating function

$$B_d^*(z) = \sum_{n=2}^{\infty} b_*(n,d)\frac{z^n}{n!} = e^z A_d(z), \tag{36}$$

where $b_*(n,d)$ is defined from (15), we find that

$$B_d^*(z) = B_{d-1}^*(pz)e^{qz} \tag{37}$$

or, since $A_d(z) = B_d^*(z)e^{-z}$,

$$A_d(z) = A_{d-1}(pz). \tag{38}$$

This can be solved by iteration to yield

$$A_d(z) = A_0(p^d z). \tag{39}$$

Then setting

$$\mathcal{G}_*(z) = \sum_{n=2}^{\infty} b(n,\infty)\frac{z^n}{n!} \tag{40}$$

and noting that

$$\sum_{n=1}^{\infty} b_*(n+1,0)\frac{z^n}{n!} = \frac{d}{dz}B_0^*(z), \tag{41}$$

(16) leads to

$$\frac{d}{dz}B_0^*(z) - B_1^*(z) = \mathcal{G}_*(z) - \mathcal{G}_*'(z). \tag{42}$$

If $\mathcal{G}_*(z) = e^z\tilde{\mathcal{G}}(z)$, from the integral representation in (9) we conclude that the Mellin transform of $\tilde{\mathcal{G}}(z)$ is

$$\int_0^\infty \tilde{\mathcal{G}}(z)z^{s-1}dz = \frac{\Gamma(s+1)}{1-p^{-s}-q^{-s}}, \tag{43}$$

Using (37), (39), and the definitions of $A_d(\cdot)$ and $\tilde{\mathcal{G}}(\cdot)$, (42) becomes

$$A_0'(z) + A_0(z) - A_0(pz) = -\tilde{\mathcal{G}}'(z). \tag{44}$$

We introduce the Mellin transform of $A_0(z)$

$$\mathcal{M}(s) = \int_0^\infty A_0(z)z^{s-1}dz \tag{45}$$

and use (44) to obtain the functional equation

$$-(s-1)\mathcal{M}(s-1) + (1-p^{-s})\mathcal{M}(s) = \frac{(s-1)\Gamma(s)}{1-p^{1-s}-q^{1-s}}. \tag{46}$$

Next we set

$$\mathcal{M}(s) = \Gamma(s)\mathcal{N}(s) \tag{47}$$

with which (46) becomes

$$-\mathcal{N}(s-1) + (1-p^{-s})\mathcal{N}(s) = \frac{s-1}{1-p^{1-s}-q^{1-s}}. \tag{48}$$

To solve (48) we let

$$\mathcal{N}(s) = \prod_{k=0}^{\infty} \left[\frac{1 - p^{k+2}}{1 - p^{k-s}} \right] \mathcal{N}_1(s) \tag{49}$$

and then (48) becomes

$$\mathcal{N}_1(s) - \mathcal{N}_1(s-1) = \prod_{k=1}^{\infty} \left[\frac{1 - p^{k-s}}{1 - p^{k+1}} \right] \frac{s-1}{1 - p^{1-s} - q^{1-s}}. \tag{50}$$

Now, for $s \to -\infty$ the right side of (50) behaves as $(s-1)\prod_{k=1}^{\infty}(1-p^{k+1})^{-1}$, with an exponentially small error. Letting

$$\mathcal{N}_1(s) = \frac{s(s-1)}{2} \prod_{k=1}^{\infty} \left(\frac{1}{1 - p^{k+1}} \right) + \mathcal{N}_2(s) \tag{51}$$

the equation for $\mathcal{N}_2(\cdot)$ becomes

$$\mathcal{N}_2(s) - \mathcal{N}_2(s-1) = \frac{s-1}{\prod_{k=1}^{\infty}(1 - p^{k+1})} \left[\frac{1}{1 - p^{1-s} - q^{1-s}} \prod_{k=1}^{\infty} (1 - p^{k-s}) - 1 \right] \tag{52}$$

whose right hand side is, unlike that of (50), exponentially small for $s \to -\infty$. The solution to (52) is

$$\mathcal{N}_2(s) = \mathcal{N}_2(-\infty) + \sum_{i=0}^{\infty} \left[\frac{\prod_{k=1}^{\infty}(1 - p^{k-s+i})}{1 - p^{1+i-s} - q^{1+i-s}} - 1 \right] \frac{s-1-i}{\prod_{k=1}^{\infty}(1 - p^{k+1})}. \tag{53}$$

Note that $\mathcal{N}_2(-\infty)$ exists, since the summand in (53) decays uniformly exponentially in i as $s \to -\infty$.

From (36) we see that $A_d(z) = O(z^2)$ as $z \to 0$ so that $\mathcal{M}(s)$ in (45) must be analytic at $s = -1$. From (47) we then conclude that $\mathcal{N}(-1) = 0$. From (49) we have $\mathcal{N}_1(-1) = 0$ and from (51) and (53) we thus obtain an expression for $\mathcal{N}_2(-\infty)$:

$$\mathcal{N}_2(-\infty) \prod_{k=1}^{\infty} (1 - p^{k+1}) + 1 - \sum_{i=0}^{\infty} (i+2) \left[\frac{\prod_{k=1}^{\infty}(1 - p^{k+i+1})}{1 - p^{2+i} - q^{2+i}} - 1 \right] = 0. \tag{54}$$

We have thus obtained the final expression for $\mathcal{M}(s)$ in (47) as

$$\mathcal{M}(s) = \frac{\Gamma(s)}{\prod_{L=0}^{\infty}(1 - p^{L-s})} \left(\frac{s(s-1)}{2} + \beta + \sum_{i=0}^{\infty} (s-i-1) \left[\frac{\prod_{k=1}^{\infty}(1 - p^{k-s+i})}{1 - p^{1+i-s} - q^{1+i-s}} - 1 \right] \right), \tag{55}$$

where

$$\beta = \mathcal{N}_2(-\infty) \prod_{k=1}^{\infty} (1 - p^{k+1})$$

can be computed from (54). Inverting the transforms in (36) and (45) we obtain

$$b_*(n,d) = \frac{n!}{2\pi i} \oint \frac{e^z}{z^{n+1}} \left[\frac{1}{2\pi i} \int_{Br} (p^d z)^{-s} \mathcal{M}(s) ds \right] dz. \tag{56}$$

The final step is to expand $b_*(n,d)$ ($\sim \tilde{b}(n,d)$) for $n \to \infty$ with d fixed. The integral over z can be asymptotically evaluated by a standard depoissonization argument, which corresponds to replacing z by n in the inner s-integral. The function $\mathcal{M}(s)$ in (55) has a triple pole at $s = 0$, and there are other double poles on the imaginary s-axis if $1 - p^{1-s} - q^{1-s}$ has zeros there, which occurs only if $\log p / \log q$ is rational, say r/t where r and t are integers [8] (cf. also [3, 19]). First we compute the contribution from $s = 0$. Using the expansion $\Gamma(s) = [1 - \gamma s + O(s^2)]/s$ as $s \to 0$, with γ being the Euler constant, (55) becomes

$$\begin{aligned} \mathcal{M}(s) &= \frac{1}{s}[1 - \gamma s + O(s^2)](1 - p^{-s})^{-1} \prod_{L=1}^{\infty} (1 - p^{L-s})^{-1} \\ &\quad \times \left(\frac{s-1}{1 - p^{1-s} - q^{1-s}} \prod_{k=1}^{\infty} (1 - p^{k-s}) - (s-1) + \frac{s(s-1)}{2} + \beta \right. \\ &\quad \left. + \sum_{i=1}^{\infty} (s - i - 1) \left[\frac{\prod_{k=1}^{\infty} (1 - p^{k-s+i})}{1 - p^{1+i-s} - q^{1+i-s}} - 1 \right] \right). \qquad (57) \end{aligned}$$

Now

$$1 - p^{-s} = s \log p - \frac{1}{2} s^2 (\log p)^2 + O(s^3)$$

and

$$1 - p^{1-s} - q^{1-s} = -hs - \frac{h_2}{2} s^2 + O(s^3).$$

Also, using the expression in (54) to compute $\beta + 1$ the expansion of (57) for $s \to 0$ becomes

$$\begin{aligned} \mathcal{M}(s) &= \frac{1}{s^3} \frac{1 - \gamma s}{\log p} \left[1 + \frac{s}{2} \log p + O(s^2) \right] \left\{ \frac{1-s}{h} \left[1 - \frac{h_2}{2h} s + O(s^2) \right] + O(s^2) \right\} \\ &= \frac{1}{s^3} \frac{1}{h \log p} + \frac{1}{s^2} \left[-\frac{\gamma}{h \log p} - \frac{1}{h \log p} \left(1 + \frac{h_2}{2h} \right) + \frac{1}{2h} \right] + O\left(\frac{1}{s} \right). \qquad (58) \end{aligned}$$

It follows that the integrand $p^{-ds} z^{-s} \mathcal{M}(s)$ in (56) has the residue

$$\operatorname{Res}_{s=0} \left\{ p^{-ds} z^{-s} \mathcal{M}(s) \right\} = \frac{1}{2} \frac{\log^2 z}{h \log p} + \frac{d}{h} \log z + \log z \left[\frac{1}{\log p} \left(\frac{\gamma + 1}{h} + \frac{h_2}{2h^2} \right) - \frac{1}{2h} \right] + O(1) \qquad (59)$$

where the $O(1)$ refers to terms that are $O(1)$ for $z \to \infty$, and these can be evaluated by explicitly computing the $O(s^{-1})$ term(s) in (58). Then the expansion of $\tilde{b}(n,d) \sim b_*(n,d)$ follows by setting $z = n$ in (59), and we have thus regained the formula in (17). If $\log p / \log q$ is rational we must also compute the contribution from the double poles along the imaginary axis at such points $p^{-s} = q^{-s} = 1$ and $p^{1-s} + q^{1-s} = 1$. These poles lead to the oscillatory terms in (17), as can be seen by computing their residues from (55).

We have thus established (17) more rigorously, though the intuitive derivation in (26)–(34) is much simpler, and more revealing of the basic asymptotic structure of the equations (13) and (14).

References

[1] Y. Choi and W. Szpankowski, Compression of Graphical Structures: Fundamental Limits, Algorithms, and Experiments, *IEEE Transaction on Information Theory*, 58(2), 620-638, 2012.

[2] L. Devroye, A Study of Trie-Like Structures Under the Density Model, *Annals of Applied Probability*, 2, 402–434, 1992.

[3] M. Drmota, *Random Trees*, Springer, New York, 2009.

[4] M. Drmota and W. Szpankowski, The Expected Profile of Digital Search Trees, *J. Combin. Theory, Ser. A*, 118, 1939–1965, 2011.

[5] P. Flajolet, X. Gourdon, and P. Dumas, Mellin Transforms and Asymptotics: Harmonic sums, *Theoretical Computer Science*, 144, 3–58, 1995.

[6] P. Flajolet, Singularity Analysis and Asymptotics of Bernoulli Sums, *Theoretical Computer Science*, 215, 371–381, 1999.

[7] P. Flajolet and R. Sedgewick, *Analytic Combinatorics*, Cambridge University Press, Cambridge, 2008.

[8] P. Flajolet, M. Roux, and B. Vallee, Digital Trees and Memoryless Sources: from Arithmetics to Analysis, AofA'10, Vienna, *Proceedings DMTCS*, 233–260, 2010.

[9] P. Jacquet, and W. Szpankowski, Analytical Depoissonization and Its Applications, *Theoretical Computer Science*, 201, 1–62, 1998.

[10] P. Jacquet and W. Szpankowski, Entropy Computations via Analytic Depoissonization, *IEEE Trans. Information Theory*, 45, 1072–1081, 1999.

[11] C. Knessl, and W. Szpankowski, Asymptotic Behavior of the Height in a Digital Search Tree and the Longest Phrase of the Lempel-Ziv Scheme, *SIAM J. Computing*, 30, 923–964, 2000.

[12] D. Knuth, *The Art of Computer Programming. Sorting and Searching*, Vol. 3, Second Edition, Addison-Wesley, Reading, MA, 1998.

[13] G. Louchard, Exact and Asymptotic Distributions in Digital and Binary Search Trees, *RAIRO Theoretical Inform. Applications*, 21, 479–495, 1987.

[14] H. Mahmoud, *Evolution of Random Search Trees*, John Wiley & Sons Inc., New York, 1992.

[15] G. Park, H.K. Hwang, P. Nicodeme, and W. Szpankowski, Profile of Tries, *SIAM J. Computing*, 8, 1821–1880, 2009.

[16] B. Pittel, Asymptotic Growth of a Class of Random Trees, *Annals of Probability*, 18, 414–427, 1985.

[17] B. Pittel, Path in a Random Digital Tree: Limiting Distributions, *Advances in Applied Probability*, 18, 139–155, 1986.

[18] W. Szpankowski, A Characterization of Digital Search Trees From the Successful Search Viewpoint, *Theoretical Computer Science*, 85, 117–134, 1991.

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