The Expected Profile of Digital Search Trees[*]

December 10, 2009

Michael Drmota[†]
Inst. für Diskrete Mathematik und Geometrie
TU Wien
A-1040 Wien,
Austria
michael.drmota@tuwien.ac.at

Wojciech Szpankowski[‡]
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Abstract

A digital search tree (DST) is a fundamental data structure on words that finds myriad of applications from the popular Lempel-Ziv'78 data compression scheme to distributed hash tables. It is a digital tree in which strings (keys, words) are stored directly in (internal) nodes. The *profile* of a DST measures the number of nodes at the same distance from the root; it is a function of the number of stored strings and the distance from the root. Most parameters of DST (e.g., depth, height, fill-up) can be expressed in terms of the profile. We make here the first step towards deciphering an asymptotic behavior of the DST profile, a long standing open problem in the analysis of algorithms and combinatorics. Throughout we assume that strings stored in DST are generated by a memoryless source. We present a precise analysis of the average profile described by a sophisticated recurrence equation that we solve by analytic methods. This analysis is surprisingly demanding but once it is carried out it reveals unusually intriguing and interesting behavior. The average profile undergoes several phase transitions when moving from the root to the longest path: at first it resembles a full tree until it abruptly starts growing polynomially and oscillating in this range. These results are derived by methods of analytic combinatorics such as generating functions, Mellin transform, Poissonization and de-Poissonization, the saddle-point method, singularity analysis and uniform asymptotic analysis.

Index Terms: Digital search trees, tree profiles, analytic combinatorics, analysis of algorithms, generating functions, Poissonization, Mellin transform.

[*] The material in this paper was presented in part at the *SODA*, New York, 2009.

[*] The work of this author was supported by the Austrian Science Foundation FWF Grant No. S9604.

[†] This work was supported in part by the NSF Grants DMS-0800568 and CCF-0830140, NSA Grant H98230-08-1-0092, and the AFOSR Grant FA8655-08-1-3018.

1 Introduction

Digital trees are fundamental data structures on words [12, 22, 24]. Among them *tries* and *digital search trees* stand out due to myriad of applications ranging from data compression to distributed hash tables [12, 16, 17, 22, 24]. In a digital search trees, the subject of this paper, strings are directly stored in nodes. More precisely, the root contains the first string (or an empty string), and the next string occupies the right or the left child of the root depending on whether its first symbol is “0” or “1”. The remaining strings are stored in available nodes which are directly attached to nodes already existing in the tree (cf. Figure 1). The search for an available node follows the prefix structure of a new string [12, 16, 22].

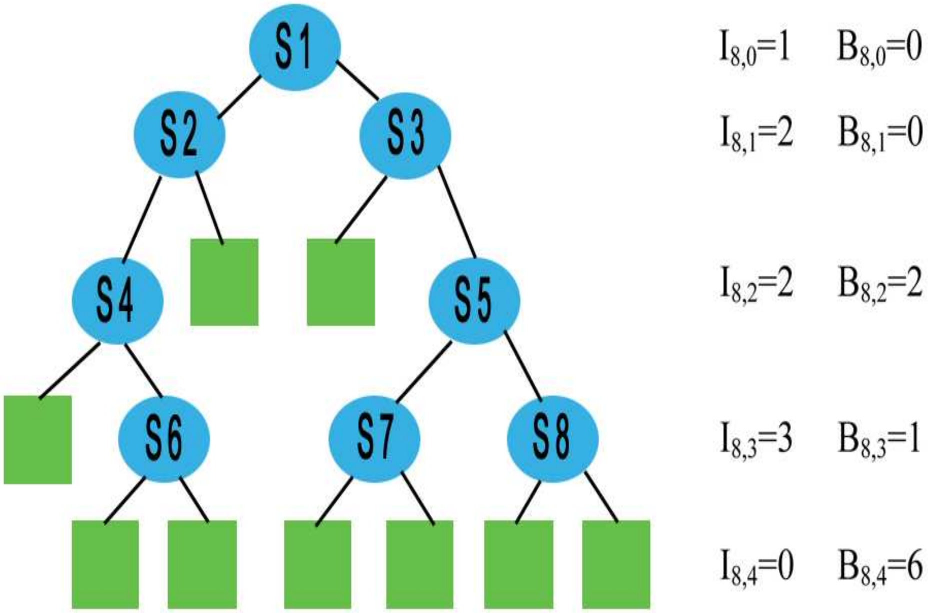

Figure 1: A digital search tree built on eight strings $s_1, \ldots, s_8$ (i.e., $s_1 = 0\ldots$, $s_2 = 1\ldots$, $s_3 = 01\ldots$, $s_4 = 11\ldots$, etc.) with internal (ovals) and external (squares) nodes, and its profiles.

In this paper, we are concerned with probabilistic properties of the *profile* defined as the number of nodes with the same distance from the root. Throughout the paper, we write $I_{n,k}$ for the number of nodes at level k when n strings are stored, and $B_{n,k}$ for the number of external nodes at level k. A digital search tree with n internal nodes is “completed” with $n + 1$ external nodes, as shown in Figure 1. These nodes can be seen as those positions where the next item can be stored. The resulting tree is then a complete binary tree with the external nodes as leaves. We study the profile built over n binary strings generated by a memoryless source, that is, we assume each string is a binary i.i.d. sequence with p being the probability of a “1” ($0 < p < 1$); we also use $q := 1 - p \leq p$. This simple model may seem too idealized for practical purposes, however, the typical behaviors under such a model often hold under more general models such as Markovian or dynamical sources, although the technicalities are usually more involved (cf. [6, 13, 24]).

The motivation of studying the profiles is multifold. First, digital search trees are used in various applications ranging from data compression (e.g., Lempel-Ziv’78 data compression scheme[1] [7]), to telecommunication (e.g., conflict resolution algorithms [24]), to partial matching of multidimensional data [22], to distributed hash tables [17]). Second, the profile is a fine shape measure closely connected to many other cost measures as further discussed below. Third, not only the analytic problems are mathematically challenging, but the diverse new phenomena they exhibit are highly interesting and unusual.

As mentioned above, several DST parameters can be expressed in terms of the internal profile:

(i) *height*: the length of the longest path from the root is $H_n = \max\{j : I_{n,j} > 0\}$;
(ii) *fill-up (or saturation) level*: the largest full level, or $F_n = \max\{j : I_{n,j} = 2^j\}$;
(iii) *depth*: the distance from the root to a randomly selected node; its distribution is given by the expected profile divided by n, [15];
(iv) *total path length*: the sum of distances between nodes and the root, or equivalently $L_n = \sum_j jI_{n,j}$.

The major difference between most previous study and the current paper is that we are dealing with asymptotics of a bivariate recurrence never addressed before in literature [4, 16, 12, 24]. We study here the following recurrence

$$x_{n+1,k+1} = \sum_{0\le j\le n} \binom{n}{j} p^j (1-p)^{n-j} \left(x_{j,k-1} + x_{n-j,k-1}\right)$$

with suitable initial conditions. We solve it asymptotically for a wide range of n and $k \le n$. This is our main contribution. We accomplish it by first considering the Poisson generating function $\Delta_k(z) := e^{-z}\sum_n x_{n,k} z^n/n!$ that satisfies the following functional-differential equation

$$\Delta'_{k+1}(z) + \Delta_{k+1}(z) = \Delta_k(pz) + \Delta_k(qz), \tag{1}$$

with a suitable $\Delta_0(z)$. This equation is still not ready for analytic combinatorics, therefore, one applies the Mellin transform, and some additional transformations leading to the following functional-recurrence equation

$$F_{k+1}(s) - F_{k+1}(s-1) = (p^{-s} + q^{-s})F_k(s) \tag{2}$$

for complex s. We are able to obtain an explicit solution of this complicated equation by introducing a proper functional operator. Next we find the inverse of the Mellin transform that leads us to an infinite number of saddle points, a rather unexpected situation (cf. also [19]). The final step is to invert asymptotics of the Poisson function $\Delta_k(z)$ through the so called *analytic depoissonization* [8] to recover asymptotically $x_{n,k}$. The reader is referred to [4, 24] for a detailed discussion of the above mentioned tools that belong to analytic combinatorics.

Digital trees have been intensively studied for the last thirty years [7, 10, 11, 12, 14, 15, 20, 23, 24], but not the profile. The closest related quantity is the typical depth D_n that measures the path length from the root to a randomly selected node; it is equal to ratio of the average profile to the number of nodes. Unfortunately, all estimations of the depth [12, 14, 15, 23, 21] deal only with the typical depth around most likely value, namely $k = 1/h \log n + O(1)$ where $h = -p\log p - q\log q$ is the entropy rate. External and internal profiles of tries have been studied by Park et. al [18, 19], while the profile of the digital search trees for *unbiased source*

[1] In particular, $I_{n,k}$ represents the number of phrases of length k in the Lempel-Ziv'78 built over n phrases.

(i.e., $p = q = 1/2$) has been recently obtained in [11] (cf. Section 6.3 of Knuth [12] for preliminary studies). The profile of digital search trees for a biased memoryless sources was left untouched for the last thirty years, and seems to be the most challenging problem in this area.

In this paper, we mostly analyze precisely the expected profile of the *biased* digital search tress and reveal unusually intriguing and interesting behavior. The average internal profile undergoes several phase transitions when moving from the root to the longest path. At first it resembles a full tree until it abruptly starts growing polynomially. Furthermore, the expected profile is oscillating in a range where the profile grows polynomially. These oscillations are due to an infinite number of saddle points. Knowing the expected profile for all values of k, we easily obtain (known and unknown) results for the typical depth and width. For example, we shall show an unusual Local Limit Theorem for the typical depth. Furthermore, our results are in accordance with known results on height, and fill up level.

The paper is organized as follows. We first present our main results and their consequences. We prove them in two sections: In Section 3 we only consider the symmetric DST (i.e., for unbiased memoryless sources). In Section 4 we deal with asymmetric DST. This is our main mathematical contribution, where we apply tools of analytic combinatorics such as poissonization, Mellin transform, and saddle point method to first solve the functional equations (1)–(2), and then extract asymptotics of the average profiles.

2 Main Results

In this section we present our main results. We first derive a general formula for the generating functions of the external and internal profiles. Then we discuss separately the symmetric case (i.e., unbiased memoryless source with $p = 0.5$), and the asymmetric case (biased memoryless source).

2.1 Generating Functions

Let $B_{n,k}$ resp. $I_{n,k}$ denote the (random) number of external resp. internal nodes at level k in a digital search tree built over n strings generated by a memoryless source with parameter $p \geq 1 - p := q$; see Figure 1. The probability generating function of the external profile, $P_{n,k}(u) = \mathbb{E}\, u^{B_{n,k}}$, satisfies the following recurrence relation (cf. [7])

$$P_{n+1,k}(u) = \sum_{\ell=0}^{n} \binom{n}{\ell} p^\ell q^{n-\ell} P_{n,k-1}(u) P_{n,k-1}(u). \tag{3}$$

The corresponding exponential generating function

$$G_k(x,u) = \sum_{n\geq 0} P_{n,k}(u) \frac{x^n}{n!}$$

fulfills the following functional recurrence

$$\frac{\partial}{\partial x} G_k(x,u) = G_{k-1}(px,u) G_{k-1}(qx,u), \qquad (k \geq 1), \tag{4}$$

with initial conditions $G_0(x,u) = u + e^x - 1$ and $G_k(0,u) = 1$ $(k \geq 1)$. Similarly, the corresponding generating function for the internal profile

$$\overline{G}_k(x,u) = \sum_{n\geq 0} \mathbb{E}\, u^{I_{n,k}} \frac{x^n}{n!}$$

satisfies the same recurrence relation

$$\frac{\partial}{\partial x}\overline{G}_k(x,u) = \overline{G}_{k-1}(px,u)\overline{G}_{k-1}(qx,u), \qquad (k \geq 1), \tag{5}$$

however, the initial conditions are $\overline{G}_0(x,u) = 1 + u(e^x - 1)$ and $\overline{G}_k(0,u) = 1$ $(k \geq 1)$.

We are interested in the expected profiles $\mathbb{E}\, B_{n,k}$ and $\mathbb{E}\, I_{n,k}$. By taking derivatives with respect to u and setting $u = 1$ we obtain for the exponential generating function

$$E_k(x) = \sum_{n\geq 0} \mathbb{E}\, B_{n,k}\, \frac{x^n}{n!}$$

the following functional recurrence

$$E_k'(x) = e^{qx}E_{k-1}(px) + e^{px}E_{k-1}(qx), \tag{6}$$

with initial condition $E_0(x) = 1$ and $E_k(0) = 0$ $(k \geq 1)$. The corresponding generating function for the internal profile

$$\overline{E}_k(x) = \sum_{n\geq 0} \mathbb{E}\, I_{n,k}\, \frac{x^n}{n!}$$

satisfies recurrence (6), too, however with initial conditions $\overline{E}_0(x) = e^x - 1$ and $\overline{E}_k(0) = 0$ $(k \geq 1)$. Note that (6) is equivalent to the recurrence relation

$$\mathbb{E}\, B_{n+1,k+1} = \sum_{\ell=0}^{n} \binom{n}{\ell} p^\ell q^{n-\ell} (\mathbb{E}\, B_{\ell,k} + \mathbb{E}\, B_{n-\ell,k}) \qquad (n, k \geq 0). \tag{7}$$

In this paper we analyze (7) for a wide range of n and k to present exact and asymptotic solutions. We first consider the symmetric case $(p = q)$, and then the asymmetric case.

2.2 Symmetric Case

Let us start with the symmetric case $p = q = \frac{1}{2}$. The corresponding generating functions have simpler structures. Namely,

$$\frac{\partial}{\partial x}G_k(x,u) = G_{k-1}\left(\frac{x}{2}, u\right)^2, \qquad (k \geq 1),$$

with initial conditions $G_0(x,u) = u + e^x - 1$ and $G_k(0,u) = 1$ (for $k > 0$), and

$$\frac{\partial}{\partial x}\overline{G}_k(x,u) = \overline{G}_{k-1}\left(\frac{x}{2}, u\right)^2, \qquad (k \geq 1),$$

with the initial conditions $\overline{G}_0(x,u) = 1 + u(e^x - 1)$ and $\overline{G}_k(0,u) = 1$. Thus (6) becomes

$$E_k'(x) = 2e^{x/2}E_{k-1}\left(\frac{x}{2}\right), \tag{8}$$

with $E_0(x) = 1$ and $E_k(0) = 0$ for $k \geq 1$ and for the internal profile

$$\overline{E}_k'(x) = 2e^{x/2}\overline{E}_{k-1}\left(\frac{x}{2}\right), \tag{9}$$

with $\overline{E}_{k-1}(0) = e^x - 1$ and $\overline{E}_k(0) = 0$.

In this special case, we can solve explicitly the above functional-differential equations leading to our first result.

Theorem 1 *Set* $Q_0 = 1$ *and*

$$Q_\ell = \prod_{j=1}^{\ell}\left(1 - \frac{1}{2^j}\right) \qquad (\ell > 0).$$

Then

$$E_k(x) = 2^k e^x \sum_{m=0}^{k} \frac{(-1)^m 2^{-\binom{m}{2}}}{Q_m Q_{k-m}} e^{-x2^{m-k}}, \tag{10}$$

and

$$\overline{E}_k(x) = 2^k e^x \left(1 - \sum_{m=0}^{k} \frac{(-1)^m 2^{-\binom{m+1}{2}}}{Q_m Q_{k-m}} e^{-x2^{m-k}}\right). \tag{11}$$

Furthermore,

$$\mathbb{E}\, B_{n,k} = 2^k \sum_{m=0}^{k} \frac{(-1)^m 2^{-\binom{m}{2}}}{Q_m Q_{k-m}} \left(1 - \frac{1}{2^{k-m}}\right)^n \tag{12}$$

and

$$\mathbb{E}\, I_{n,k} = 2^k - 2^k \sum_{m=0}^{k} \frac{(-1)^m 2^{-\binom{m+1}{2}}}{Q_m Q_{k-m}} \left(1 - \frac{1}{2^{k-m}}\right)^n. \tag{13}$$

for any n *and* $k \le n$.

There are several ways to prove these relations. The simplest way is to use induction. It should be noted that the explicit formula (12) for $\mathbb{E}\, B_{n,k}$ has appeared several times in the literature [14, 15, 16, 21]. Therefore, we omit here details of the proof.

In Section 3 we establish asymptotic behavior of the average profiles presented next.

Theorem 2 *We have*

$$\mathbb{E}\, B_{n,k} = 2^k F'(n2^{-k}) + F''(n2^{-k}) + O(n2^{-k}) \tag{14}$$

and

$$\mathbb{E}\, I_{n,k} = 2^k F(n2^{-k}) + F'(n2^{-k}) + O(n2^{-k}) \tag{15}$$

where

$$F(z) = 1 - \sum_{m\ge 0} \frac{(-1)^m 2^{-\binom{m+1}{2}}}{Q_\infty Q_m} e^{-z2^m} \tag{16}$$

uniformly for all $n, k \ge 1$, *where* $Q_\infty = \prod_{j\ge1}(1 - 2^{-j})$. *For* $n2^{-k} \to 0$, *we can use the following asymptotic expression for* $F(z)$

$$F^{(r)}(z) = C_r(z) 2^{-\frac{1}{2}\left(\log_2 \frac{1}{z}\right)^2 - \log_2 \frac{1}{z} \log_2 \log_2 \frac{1}{z}}, \qquad as \quad z \to 0^+,$$

where $F^{(r)}(z)$ *is the* r*th derivative of* $F(z)$ *and* $C_r(z)$ *is a function of at most polynomial growth.*

In passing we should point out that a precise asymptotic behavior of the internal profile of the symmetric DST for a wide range of $k \le n$ was also recently presented in [11].

2.3 Asymmetric Case

The asymmetric case ($p < q$) is much more involved. In particular, we cannot obtain a simple exact solution for the exponential generating function $E_k(x)$. To circumvent this problem, we apply the Poisson transform and the Mellin transform [4, 24] to find asymptotic solutions.

Let us start with the external profile. The Poisson transform of $E_k(x)$, namely

$$\Delta_k(x) = e^{-x} \sum_{n \geq 0} \mathbb{E}\, B_{n,k} \frac{x^n}{n!} = E_k(x) e^{-x}, \qquad (k \geq 0)$$

translates recurrence (6) into

$$\Delta_k(x) + \Delta'_k(x) = \Delta_{k-1}(px) + \Delta_{k-1}(qx), \qquad (k \geq 1), \tag{17}$$

with initial conditions $\Delta_0(x) = e^{-x}$ and $\Delta_k(0) = 0$ ($k \geq 1$). This can be solved using the Mellin transform discussed next.

The Mellin transform of $\Delta_k(x)$ is defined as [4, 24]

$$\Delta^*_k(s) = \int_0^\infty \Delta_k(x) x^{s-1}\, dx.$$

By induction it is easy to prove that $\Delta_k(x)$ can be represented as a finite linear combination of functions of the form $e^{-p^{\ell_1} q^{\ell_2} x}$ with $\ell_1, \ell_2 \geq 0$ and $0 \leq \ell_1 + \ell_2 \leq k$. Hence, $\Delta^*_k(s)$ exists for all s with $\Re(s) > 0$. Furthermore, $B_{n,k} = 0$ for $k > n$. Thus, $E_k(x) = O(x^k)$ for $x \to 0$ which ensures that $\Delta^*_k(s)$ actually exists for s with $\Re(s) > -k$.

Let us now express $\Delta^*_k(s)$ as

$$\Delta^*_k(s) = \Gamma(s) F_k(s),$$

where $\Gamma(s)$ is the Euler gamma function. In the above, $F_k(s)$ is a finite linear combination of functions of the form $p^{-\ell_1 s} q^{-\ell_2 s}$ with $\ell_1, \ell_2 \geq 0$ and $0 \leq \ell_1 + \ell_2 \leq k$. Thus, $F_k(s)$ can be considered an entire function. Then (17) translates into

$$F_k(s) - F_k(s-1) = (p^{-s} + q^{-s}) F_{k-1}(s) = T(s) F_{k-1}(s), \qquad (k \geq 1), \tag{18}$$

with initial condition $F_0(s) = 1$ and

$$T(s) = p^{-s} + q^{-s}. \tag{19}$$

Note that (18) does not only hold for $\Re(s) > -k$ where the Mellin transform exists. Since $F_k(s)$ analytically continues to an entire function, (18) holds for all s.

In order to find a solution of (18) we define the power series

$$f(s, w) = \sum_{k \geq 0} F_k(s) w^k.$$

Let us also introduce a function operator $\mathbf{A}$ as follows

$$\mathbf{A}[f](s) = \sum_{j \geq 0} f(s-j) T(s-j). \tag{20}$$

In the next theorem we find an explicit representation of $F_k(x)$ through the operator $\mathbf{A}$. The proof is delayed till Section 4.

Theorem 3 *The functions $F_k(s)$ are recursively given by*

$$F_k(s) = \mathbf{A}[F_{k-1}](s) - \mathbf{A}[F_{k-1}](0) \qquad (k \ge 1) \tag{21}$$

with initial function $F_0(s) = 1$ and $F_k(-\ell) = 0$ for $\ell = 0, 1, 2, \ldots, k-1$ and $k \ge 1$.

Furthermore, if we set $R_k(s) = \mathbf{A}^k[1](s)$ then we have the formal identity

$$\sum_{k\ge 0} F_k(s) w^k = \frac{\sum_{\ell \ge 0} R_\ell(s) w^\ell}{\sum_{\ell\ge 0} R_\ell(0) w^\ell}, \tag{22}$$

Remark 1 It is easy to compute $R_k(s)$ for a few small values of k. For example,

$$\begin{aligned} R_0(s) &= 1, \\ R_1(s) &= \frac{p^{-s}}{1-p} + \frac{q^{-s}}{1-q}, \\ R_2(s) &= \frac{p^{-2s}}{(1-p)(1-p^2)} + \frac{p^{-s}q^{-s}}{(1-p)(1-pq)} + \frac{p^{-s}q^{-s}}{(1-q)(1-pq)} + \frac{q^{-2s}}{(1-q)(1-q^2)}. \end{aligned}$$

In Section 4 we use the above representation to find the asymptotic behavior of the average profiles. To present it in a concise form, we need some additional notation. For a real number α with $(\log\frac{1}{p})^{-1} < \alpha < (\log\frac{1}{q})^{-1}$, let

$$\rho = \rho(\alpha) = \frac{1}{\log(p/q)} \log \frac{1 - \alpha\log(1/p)}{\alpha\log(1/q) - 1}. \tag{23}$$

Equivalently, α and ρ satisfy the equation

$$\alpha = \frac{p^{-\rho} + q^{-\rho}}{p^{-\rho}\log\frac{1}{p} + q^{-\rho}\log\frac{1}{q}}.$$

Furthermore, we set

$$\beta(\rho) = \frac{p^{-\rho}q^{-\rho}\log(p/q)^2}{(p^{-\rho}+q^{-\rho})^2}, \tag{24}$$

and we also use the abbreviation

$$\alpha_0 = \frac{2}{\log\frac{1}{p} + \log\frac{1}{q}}.$$

Our first main asymptotic result is presented next.

Theorem 4 *Let $\mathbb{E}\, B_{n,k}$ denote the expected external profile in (asymmetric) digital search trees with $0 < p < q = 1-p < 1$. If n and k are positive integers with $\frac{1}{\log\frac{1}{p}} + \varepsilon \le \frac{k}{\log n} \le \frac{1}{\log\frac{1}{q}} - \varepsilon$ (for some $\varepsilon > 0$), then uniformly*

$$\mathbb{E}\, B_{n,k} = H\left(\rho_{n,k}, \log_{p/q} p^k n\right) \frac{(p^{-\rho_{n,k}} + q^{-\rho_{n,k}})^k n^{-\rho_{n,k}}}{\sqrt{2\pi\beta(\rho_{n,k})k}} \left(1 + O\left(k^{-1/2}\right)\right), \tag{25}$$

where $\rho_{n,k} = \rho(k/\log n)$ and $H(\rho, x)$ is a non-zero periodic function with period 1 given by (53) of Section 4.

The average internal profile is slightly more complicated. Among others, it exhibits some phase transition discussed below. As before, the Poisson transform of $\mathbb{E}\, I_{n,k}$ is defined as

$$\overline{\Delta}_k(x) = e^{-x} \sum_{n\ge 0} \mathbb{E}\, I_{n,k} \frac{x^n}{n!} = E_k(x) e^{-x}, \qquad (k \ge 0)$$

which translates into

$$\overline{\Delta}_k(x) + \overline{\Delta}_k'(x) = \overline{\Delta}_{k-1}(px) + \overline{\Delta}_{k-1}(qx), \qquad (k \ge 1), \tag{26}$$

with the initial condition $\overline{\Delta}_0(x) = 1 - e^{-x}$. This initial condition shifts the existence of the Mellin transform $\overline{\Delta}_k^*(s)$ to $-k-1 < \Re(s) < 0$. Let now

$$\overline{\Delta}_k^*(s) = -\Gamma(s)\overline{F}_k(s)$$

where $\overline{F}_0(s) = 1$ and by (26) we find

$$\overline{F}_k(s) - \overline{F}_k(s-1) = T(s)\overline{F}_{k-1}(s).$$

Using now the operator $\mathbf{A}$ defined in (20), we can express $\overline{F}_k(s)$ similarly as in Theorem 3, that is,

$$\overline{F}_k(s) = \mathbf{A}[\overline{F}_{k-1}](s) - \mathbf{A}[\overline{F}_{k-1}](-1) \qquad (k \ge 1) \tag{27}$$

and

$$\sum_{k\ge 0} \overline{F}_k(s) w^k = \frac{\sum_{\ell\ge 0} R_\ell(s) w^\ell}{\sum_{\ell \ge 0} R_\ell(-1) w^\ell}. \tag{28}$$

Using this representation, in Section 4 we prove our second main asymptotic result.

Theorem 5 *Let $\mathbb{E}\, I_{n,k}$ denote the expected internal profile in (asymmetric) digital search trees with $0 < p < q = 1-p < 1$. Let k and n be positive integers such that $k/\log n$ satisfies $(\log \frac{1}{p})^{-1} < k/\log n < (\log \frac{1}{q})^{-1}$. Then the following assertions hold:*

1. *If $\frac{1}{\log \frac{1}{p}} + \varepsilon \le \frac{k}{\log n} \le \alpha_0 - \varepsilon$ (for some $\varepsilon > 0$), then uniformly*

$$\mathbb{E}\, I_{n,k} = 2^k - \overline{H}\left(\rho_{n,k}, \log_{p/q} p^k n\right) \frac{(p^{-\rho_{n,k}} + q^{-\rho_{n,k}})^k n^{-\rho_{n,k}}}{\sqrt{2\pi\beta(\rho_{n,k})k}} \left(1 + O\left(k^{-1/2}\right)\right),$$

where $\overline{H}(\rho, x)$ is a non-zero periodic function with period 1 (see Section 4 for more details).

2. *If $k = \alpha_0 \left(\log n + \xi\sqrt{\alpha_0 \beta(0) \log n}\right)$, where $\xi = o((\log n)^{\frac{1}{6}})$, then*

$$\mathbb{E}\, I_{n,k} = 2^k \Phi(-\xi) \left(1 + O\left(\frac{1 + |\xi|^3}{\sqrt{\log n}}\right)\right)$$

where Φ is the normal distribution function.

3. *If $\alpha_0 + \varepsilon \le \frac{k}{\log n} \le \frac{1}{\log \frac{1}{q}} - \varepsilon$ (for some $\varepsilon > 0$) then uniformly*

$$\mathbb{E}\, I_{n,k} = \overline{H}\left(\rho_{n,k}, \log_{p/q} p^k n\right) \frac{(p^{-\rho_{n,k}} + q^{-\rho_{n,k}})^k n^{-\rho_{n,k}}}{\sqrt{2\pi\beta(\rho_{n,k})k}} \left(1 + O\left(k^{-1/2}\right)\right)$$

with the same function $\overline{H}(\rho, x)$ as in 1.

Finally, we point out that if we set $\alpha = k/\log n$, then we can rewrite

$$(p^{-\rho} + q^{-\rho})^k n^{-\rho} = n^{\alpha \log(p^{-\rho}+q^{-\rho})-\rho}.$$

Thus, for $\alpha_0 < \alpha < \alpha_2$ the behavior of $\mathbb{E}\, B_{n,k}$ and $\mathbb{E}\, I_{n,k}$ is governed by a power of n depending on the ratio $\alpha = k/\log n$. The maximum exponent is obtained for

$$\alpha = \frac{1}{h} = \frac{1}{p \log \frac{1}{p} + q \log \frac{1}{q}},$$

where $h = p\log\frac{1}{p} + q\log\frac{1}{q}$ denotes the entropy of the Bernoulli source. Actually, the expected number of nodes at level $k = \frac{1}{h}\log n$ is of order $n/\sqrt{\log n}$.

2.4 Some Consequences

In this section we briefly present some consequences of our main findings. We start with the typical depth. Let D_n denote the depth of a random node in a digital search tree with n nodes. Then the distribution of D_n is related to the external profile by [12, 23]

$$\mathbb{P}\{D_n = k\} = \frac{\mathbb{E}\, I_{n,k}}{n}.$$

Hence, a direct application of Theorem 5 provides an unusual local limit theorem.

Theorem 6 *Let D_n denote the depth of a random node in a binary random digital search tree with $0 < p < q = 1-p < 1$. Then*

$$\mathbb{P}\{D_n = k\} = \frac{\overline{H}\left(-1, \log_{p/q} p^k n\right)}{\sqrt{2\pi(h_2 - h^2)/h^3 \, \log n}} \exp\left(-\frac{\left(k - \frac{1}{h}\log n\right)^2}{2(h_2-h^2)/h^3 \, \log n}\right) \times \left(1 + O\left(\frac{1}{\sqrt{\log n}} + \frac{\left|k - \frac{1}{h}\log n\right|^3}{(\log n)^2}\right)\right)$$

uniformly for k and n with $\left|k - \frac{1}{h}\log n\right| = o\left((\log n)^{2/3}\right)$ where $h_2 = p(\log\frac{1}{p})^2 + q(\log\frac{1}{q})^2$.

The unusualness of this result is the periodic factor $\overline{H}(\cdot,\cdot)$ in the local limit theorem. Although the depth D_n follows a central limit theorem (see [15]) it does not obey the corresponding local central limit theorem (compare also with [19]).

As a further corollary to the above finding, we observe that the *width* W_n (defined as $\max_k I_{n,k}$) satisfies

$$\mathbb{E}\, W_n \ge \max_k \mathbb{E}\, I_{n,k} = \Omega\left(\frac{n}{\log n}\right).$$

In order to obtain a corresponding upper bound (one expects that the order of magnitude of the lower bound is the correct one) we would need some information about the second moment $\mathbb{E}\, I_{n,k}^2$, compare with [2].

Remark 2 Other parameters of interest are the height $H_n = \max\{k : \; I_{n,k} > 0\}$ and the fillup level $F_n = \max\{k : \; I_{n,k} = 2^k\}$. It is well known (see [20, 1]) that

$$\frac{H_n}{\log n} \to \frac{1}{\log p^{-1}}, \quad \text{in probability},$$

and

$$\frac{F_n}{\log n} \to \frac{1}{\log q^{-1}}, \quad \text{in probability.}$$

This is completely in accordance with our findings. Theorem 4 only works in the *interesting range* $(\log \frac{1}{p})^{-1} < \alpha = k/\log n < (\log \frac{1}{q})^{-1}$ where we either have $2^k - \mathbb{E}\, I_{n,k} \to \infty$ resp. $\mathbb{E}\, I_{n,k} \to \infty$. However, it is a common phenomenon that fillup level and height *occur*, where $\mathbb{E}\, I_{n,k} = 2^k - O(1)$ resp. where $\mathbb{E}\, I_{n,k} = O(1)$. By extrapolating the asymptotic expansions, presented in Theorem 4, to its boundaries $\alpha = (\log \frac{1}{p})^{-1}$ resp. $\alpha = (\log \frac{1}{q})^{-1}$ one can expect to prove this kind of behavior (cf. [19] for tries). In order to make this precise we would have to discuss the asymptotic behavior of $F_k(s)$ for $s \to \infty$ and $s \to -\infty$, which we omit in this paper (cf. [10] for the symmetric case).

3 Analysis: Symmetric Case

In the symmetric case, we only discuss the asymptotic analysis of the profile. We prove here (14) and (15) of Theorem 2 that we repeat below:

$$\mathbb{E}\, B_{n,k} = 2^k F'(n2^{-k}) + F''(n2^{-k}) + O(n2^{-k})$$

and

$$\mathbb{E}\, I_{n,k} = 2^k F(n2^{-k}) + F'(n2^{-k}) + O(n2^{-k}),$$

where (cf. (16))

$$F(z) = 1 - \sum_{m\ge 0} \frac{(-1)^m 2^{-\binom{m+1}{2}}}{Q_\infty Q_m} e^{-z2^m}.$$

It is easy to see that

$$F(z) = 1 - \frac{1}{Q_\infty} e^{-z} + O(e^{-2z}) \qquad (z \to \infty).$$

However, for the asymptotics of $\mathbb{E}\, B_{n,k}$ and $\mathbb{E}\, I_{n,k}$ we need the behavior of $F(z)$ for $z \to 0+$ which is much more involved and will be presented in Lemma 2. Interestingly we need on the following identity.[2]

Lemma 1 *Suppose that $|q| < 1$ and let*

$$(a;q)_\infty = \prod_{j=0}^{\infty} (1 - aq^j) \qquad \textit{and} \qquad (a;q)_k = \frac{(a;q)_\infty}{(aq^k;q)_\infty},$$

be the usual q-Pochhammer notation for q-rising factorials. Then

$$\sum_{k=0}^{\infty} \frac{(-1)^k q^{\binom{k}{2}}}{(q;q)_k (1 - cq^{k-1})} q^{2k} = \frac{q}{c} \frac{1}{1 - c/q} \frac{(q;q)_\infty}{(c;q)_\infty} - \frac{q}{c} (q;q)_\infty. \tag{29}$$

[2] We are grateful to Michael Schlosser (University of Vienna) who proposed a simple proof of Lemma 1.

Proof: We first recall the following ${}_1\phi_1$ summation formula [5, Appendix (II.5)]:

$$\sum_{k=0}^{\infty} \frac{(a;q)_k (-1)^k q^{\binom{k}{2}}}{(q;q)_k (c;q)_k} \left(\frac{c}{a}\right)^k = \frac{(c/a;q)_\infty}{(c;q)_\infty} \tag{30}$$

that we will apply twice. We consider the left hand side of (29) and replace one factor q^k by $\frac{q}{c} - \frac{q}{c}(1-cq^{k-1})$ that leads to

$$\begin{aligned}\sum_{k=0}^{\infty} \frac{(-1)^k q^{\binom{k}{2}}}{(q;q)_k (1-cq^{k-1})} q^{2k} &= \sum_{k=0}^{\infty} \frac{(-1)^k q^{\binom{k}{2}}}{(q;q)_k (1-cq^{k-1})} q^k \left[\frac{q}{c} - \frac{q}{c}(1-cq^{k-1})\right] \\ &= \frac{q}{c}\frac{1}{1-c/q} \sum_{k=0}^{\infty} \frac{(c/q;q)_k (-1)^k q^{\binom{k}{2}}}{(q;q)_k (c;q)_k} q^k - \frac{q}{c} \sum_{k=0}^{\infty} \frac{(-1)^k q^{\binom{k}{2}}}{(q;q)_k} q^k.\end{aligned}$$

The first sum on the right hand side can be simplified using (30) by setting $a = c/q$. For the second sum we can also apply (30) with $a = c/q$ but considering the limit $c \to 0$. This proves (29). ■

If we apply (29) with the special values $q = \frac{1}{2}$ and $c = -\frac{s}{2}$ we obtain (after some elementary calculations)

$$\frac{1}{s} \prod_{j=0}^{\infty} \frac{1}{1+s2^{-j}} = \frac{1}{s} - \prod_{j=1}^{\infty} \frac{1}{1-2^{-j}} \sum_{m=0}^{\infty} \frac{(-1)^m \, 2^{-\binom{m+1}{2}}}{(s+2^m)\prod_{j=1}^{m}(1-2^{-j})}. \tag{31}$$

This identity is now used in the proof of Lemma 2 and in an asymptotic expansion of $F(z)$ and its derivatives.

Lemma 2 *The Laplace transform $L(s) = \int_0^\infty F(z) e^{-sz}\, dz$ is given by*

$$L(s) = \frac{1}{s} \prod_{j=0}^{\infty} \frac{1}{1+s2^{-j}}. \tag{32}$$

Furthermore, for any fixed $r \ge 0$ the r-th derivative $F^{(r)}(z)$ is asymptotically equivalent to

$$F^{(r)}(z) = C_r(z) 2^{-\frac{1}{2}\left(\log_2 \frac{1}{z}\right)^2 - \log_2 \frac{1}{z} \log_2 \log_2 \frac{1}{z}}, \qquad z \to 0^+ \tag{33}$$

where $C_r(z)$ is a function of at most polynomial growth for $z \to 0+$. In particular $\lim_{z\to 0+} F(z) = 0$.

Proof: Since $F(z)$ is bounded for $z \ge 0$, the Laplace transform $L(s)$ exists for $\Re(s) > 0$ and is given by

$$L(s) = \int_0^\infty F(z) e^{-sz}\, dz = \frac{1}{s} - \sum_{m \ge 0} \frac{(-1)^m 2^{-\binom{m+1}{2}}}{Q_\infty \, Q_m \, (s+2^m)}$$

Applying (29) we prove (32) of $L(s)$.

Denoting

$$Q(x) = \prod_{j=1}^{\infty} \left(1 - \frac{x}{2^j}\right)$$

we also have

$$L(s) = \frac{1}{s\,Q(-2s)}.$$

Note that the Laplace transforms of the derivatives $F^{(r)}(z)$ are given by $s^r L(s)$.

In order to obtain the asymptotic expansion (33) we use the integral representation for the inverse Laplace transform

$$F^{(r)}(z) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} s^r L(s) e^{sz}\,ds,$$

where c is an arbitrary positive number (which we choose in the sequel). The idea is to use a saddle point approximation of the integrand. For this purpose we need an asymptotic formula for $Q(-2x)$ for $x \to \infty$:

$$Q(-2x) = \exp\left(\frac{(\log x)^2}{2\log 2} + \frac{\log x}{2} + \frac{\pi^2}{6\log 2} - \frac{1}{6} + \frac{1}{4}\log 2 + \Psi(\log_2(x)) + O(1/x)\right), \qquad (34)$$

where $\Psi(x)$ is a differentiable periodic function with period 1. This follows from the Mellin transform applied to the logarithms. Indeed, for $-1 < \Re(u) < 0$, using the "harmonic sum formula" [3, 24] we find

$$M(u) = \int_0^\infty \log Q(-2x) x^{u-1}\,dx = \frac{1}{1-2^u} R(u)$$

with

$$R(u) = \int_0^\infty \log(1+x) x^{u-1}\,dx = \frac{\pi}{u\sin(\pi u)}$$

The inverse Mellin transform yields

$$Q(-2x) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} M(u) x^{-u}\,du$$

with $-1 < c < 0$. By shifting the line of integration to the right and collecting the *contributions* from the triple pole at $u = 0$ and the single poles at $u = ki/\log 2$ ($k \in \mathbb{Z}\setminus\{0\}$), which constitute the periodic function Ψ, we obtain (34) (compare with [3]). It is also easy to extend the asymptotic relation (34) to the complex plane $|\arg(x)| \le \delta$ and $|x| \to \infty$, where δ is a small positive number.

Finally, we evaluate the r-th derivative $F^{(r)}(z)$ asymptotically. Let $0 < z < 1$ be given. We will compute the integral

$$F^{(r)}(z) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \frac{s^{r-1}}{Q(-2s)} e^{sz}\,ds,$$

where

$$c = c(z) = \frac{\log(1/z)}{z\log 2} = \frac{\log_2(1/z)}{z}$$

as the (approximate) saddle point of the function

$$s \mapsto \exp\left(sz - \frac{(\log x)^2}{2\log(2)}\right).$$

Hence, by a standard saddle point method we obtain, after some algebra

$$F^{(r)}(z) \sim C' e^{-\Psi(\log_2 c(z))} c(z)^{r+\frac{1}{2}} (\log c(z))^{-\frac{3}{2}} e^{\log_2 \frac{1}{z} - \frac{\log 2}{2} (\log_2 c(z))^2},$$

where C' is a constant. This completes the proof of Lemma 2 (see also [10]). ∎

Remark 3 We note that the relation $I_{n,k} = \sum_{j \ge 1} 2^{-j} B_{n,k+j}$ is reflected by the functional equation

$$F(z) = \sum_{j \ge 1} F'(z 2^{-j})$$

which is equivalent to the relation

$$\overline{L}(s) = s \sum_{j \ge 1} 2^j \overline{L}(s 2^j), \tag{35}$$

where $\overline{L}(s) = sL(s)$ denotes the Laplace transform of $F'(z)$. Interestingly the equation (35) is deduced easily from the identity

$$\sum_{j \ge 1} \frac{2^j}{(1+2s)(1+2^2 s)\cdots(1+2^j s)} = \frac{1}{s},$$

that follows from (30) by setting $a = q = \frac{1}{2}$, $c = -1/(4s)$, applying the index shift $k \mapsto j-1$ and doing some elementary calculations.

Proof of Theorem 2: In order to Prove Theorem 2 we use the the explicit representations (11) and (12) of Theorem 1 and approximate the leading terms by $2^k F'(n2^{-k})$ resp. by $2^k F(n2^{-k})$.

Let us concentrate on the external profile. We repeat here equation (12)

$$\mathbb{E}\, B_{n,k} = 2^k \sum_{m=0}^{k} \frac{(-1)^m 2^{-\binom{m}{2}}}{Q_m Q_{k-m}} \left(1 - \frac{1}{2^{k-m}}\right)^n .$$

We first show that the terms for $m > k/3$ can be neglected

$$\left| \sum_{m > k/3} \frac{(-1)^m 2^{-\binom{m}{2}}}{Q_m Q_{k-m}} \left(1 - 2^{m-k}\right)^n \right| = O\left(\sum_{m > k/3} 2^{-\binom{m}{2}} \right) = O\left(2^{-\binom{\lfloor k/3 \rfloor}{2}} \right).$$

In a next step we use the approximation (for $m \le k/3$)

$$\left(1 - 2^{m-k}\right)^n = e^{-n 2^{m-k}} \left(1 + O\left(\frac{n}{4^{k-m}}\right)\right) = e^{-n2^{m-k}} + O\left(\frac{n}{4^{k-m}}\right),$$

and obtain an error term of the form

$$\sum_{m \le k/3} \frac{(-1)^m 2^{-\binom{m}{2}}}{Q_m Q_{k-m}} \cdot O\left(\frac{n}{4^{k-m}}\right) = O\left(\frac{n}{4^k}\right).$$

Finally, we approximate the ratio

$$\frac{Q_\infty}{Q_m} = 1 - \frac{1}{2^{k-m}} + O\left(\frac{1}{4^{k-m}}\right)$$

leading to even smaller error term $O(4^{-k})$. Summing up we arrive at

$$\begin{aligned}
\mathbb{E}\, B_{n,k} &= 2^k \sum_{m=0}^{k} \frac{(-1)^m 2^{-\binom{m}{2}}}{Q_m Q_{k-m}} \left(1 - 2^{m-k}\right)^n \\
&= 2^k \sum_{m\le k/3} \frac{(-1)^m 2^{-\binom{m}{2}}}{Q_\infty Q_m} \left(1 - \frac{1}{2^{k-m}}\right) e^{-n2^{m-k}} + O\left(2^k \frac{n}{4^k}\right) + O\left(2^k 2^{-\binom{\lfloor k/3\rfloor}{2}}\right) \\
&= 2^k \sum_{m=0}^{\infty} \frac{(-1)^m 2^{-\binom{m}{2}}}{Q_\infty Q_m} \left(1 - \frac{2^m}{2^k}\right) e^{-n2^{m-k}} + O\left(\frac{n}{2^k}\right) \\
&= 2^k F'(n2^{-k}) + F''(n2^{-k}) + O(n2^{-k}).
\end{aligned}$$

This completes the proof of (14). The proof of (15) is exactly the same. ■

Remark 4 These expansions are valid only if $n2^{-k} \le 2^k$, that is, for $k \ge \frac{1}{2}\log_2 n$. However, for small k there are no interesting phenomena. The range of interest is (cf. [10])

$$\log_2 n - \log\log n \le k \le \log_2 n + \sqrt{2\log_2 n},$$

and this range is covered by Theorem 2. Nevertheless, with slightly more care it is easy to obtain more precise expansions, e.g.

$$\mathbb{E}\, B_{n,k} = 2^k F'(n2^{-k}) + F''(n2^{-k}) - \frac{n}{2^{k+1}} F'''(n2^{-k}) + O(n4^{-k}) + O(2^{-k}).$$

4 Analysis: Asymmetric Case

In this section, we return to the asymmetric case ($p \ne q$). We first derive the exact representation for the Mellin transform $F_k(s)$, proving Theorem 3. Then we deal with asymptotic results establishing Theorems 4 and 5.

4.1 Proof of Theorem 3: Exact Representation

Let us recall that $\mathbf{A}$ is a functional operator is defined by

$$\mathbf{A}[f](s) = \sum_{j\ge 0} f(s-j) T(s-j),$$

where $T(s) = p^{-s} + q^{-s}$. We prove here Theorem 3, that is,

$$F_k(s) = \mathbf{A}[F_{k-1}](s) - \mathbf{A}[F_{k-1}](0) \qquad (k \ge 1) \tag{36}$$

where $F_0(s) = 1$, and

$$\sum_{k\ge 0} F_k(s) w^k = \frac{\sum_{\ell\ge 0} R_\ell(s) w^\ell}{\sum_{\ell\ge 0} R_\ell(0) w^\ell}. \tag{37}$$

where $R_k(s) = \mathbf{A}^k[1](s)$; also $F_k(-\ell) = 0$ for $\ell = 0, 1, 2, \ldots, k-1$.

Proof Theorem 3: Set $\tilde{F}_0(s) = 1$ and recursively

$$\tilde{F}_k(s) = \mathbf{A}[\tilde{F}_{k-1}](s) - \mathbf{A}[\tilde{F}_{k-1}](0) \qquad (k \geq 1).$$

It is easy to see that $\tilde{F}_k(s)$ is a well defined entire function. In particular it follows that $\tilde{F}_k(s)$ is (as it is for $F_k(s)$) a finite linear combination of a function of the form $p^{-\ell_1 s} q^{-\ell_2 s}$ with $\ell_1, \ell_2 \geq 0$ and $\ell_1 + \ell_2 \leq k$. Furthermore, by definition these functions satisfy $\tilde{F}_k(0) = 0$ (for $k \geq 1$) and by (36) fulfill the relation

$$\tilde{F}_k(s) - \tilde{F}_k(s-1) = T(s)\tilde{F}_{k-1}(s)$$

for $k \geq 0$ and all s.

Now we can proceed by induction to show that $F_k(s) = \tilde{F}_k(s)$. By definition we have $F_0(s) = \tilde{F}_0(s)$. Now suppose that $F_k(s) = \tilde{F}_k(s)$ holds for some $k \geq 0$. Then it follows that $F_{k+1}(s) = \tilde{F}_{k+1}(s) + G(s)$, where $G(s)$ satisfies

$$G(0) = 0 \quad \text{and} \quad G(s) - G(s-1) = 0, \qquad (\Re(s) > -k). \tag{38}$$

By the above observations $G(s)$ has to be a finite linear combination of functions of the form $p^{-\ell_1 s} q^{-\ell_2 s}$. However, the only periodic function of this form that meets conditions (38) is the zero function. Hence, $F_{k+1}(s) = \tilde{F}_{k+1}(s)$.

Now we prove (37), which is equivalent to

$$\sum_{\ell=0}^{k} F_\ell(s) R_{k-\ell}(0) = R_k(s), \qquad (k \geq 0),$$

or

$$F_k(s) = R_k(s) - \sum_{\ell=0}^{k-1} F_\ell(s) R_{k-\ell}(0), \qquad (k \geq 0).$$

We will prove this relation by induction. Certainly, it is satisfied for $k = 0$. Now suppose that is holds for some $k \geq 0$. By (36) we also have

$$\begin{aligned}
F_{k+1}(s) &= \mathbf{A}[F_k](s) - \mathbf{A}[F_k](0) \\
&= \mathbf{A}[R_k](s) - \mathbf{A}[R_k](0) - \sum_{\ell=0}^{k-1} (\mathbf{A}[F_\ell](s) - \mathbf{A}[F_\ell](0)) R_{k-\ell}(0) \\
&= R_{k+1}(s) - R_{k+1}(0) - \sum_{\ell=0}^{k-1} F_{\ell+1}(s) R_{k-\ell}(0) \\
&= R_{k+1}(s) - \sum_{\ell=0}^{k} F_\ell(s) R_{k+1-\ell}(0).
\end{aligned}$$

Finally, since $F_k(s) = -\Delta_k^*(s)/\Gamma(s)$ is analytic for $\Re(s) > -k$ and $1/\Gamma(-\ell) = 0$, it also follows that $F_k(-\ell) = 0$ for $\ell = 0, 1, \ldots, k-1$. ■

Following the same footsteps, we prove the corresponding relations for the internal profile presented just above Theorem 5; in particular, (27) and (28).

Remark 5 The proof of (36) (and consequently that of (37)) makes use of the fact that $F_k(0) = 0$ for $k \geq 1$. However, we also have $F_k(-r) = 0$ for $k > r$. In particular, if we set $s = -r$ in (37) we find

$$\sum_{k=0}^{r} F_k(-r)w^k = \frac{\sum_{\ell \geq 0} R_k(-r)w^k}{\sum_{\ell \geq 0} R_k(0)w^k},$$

and consequently

$$\sum_{k \geq 0} F_k(s)w^k = \frac{\sum_{\ell \geq 0} R_k(s)w^k}{\sum_{\ell \geq 0} R_k(-r)w^k} \sum_{k=0}^{r} F_k(-r)w^k. \tag{39}$$

4.2 Asymptotic Analysis

We now prove Theorems 4 and 5 establishing asymptotic behavior of the average profiles. The discussion is divided into several steps: First we analyze $F_k(s)$, then we invert the Mellin transform $\Delta_k^*(s) = \Gamma(s)F_k(s)$, and finally we invert the Poisson transform $\Delta_k(x)$ to obtain asymptotics for the expected profile $\mathbb{E}\, B_{n,k}$. Finally, we comment on necessary changes to recover asymptotics of $\mathbb{E}\, I_{n,k}$.

4.2.1 Singularity Analysis of $F_k(s)$

In order to obtain asymptotic information for $F_k(s)$ we will analyze the generating function $f(s,w) = \sum_k F_k(s)w^k$ that by Theorem 3 is also given by

$$f(s,w) = \frac{g(s,w)}{g(0,w)},$$

where $g(s,w) = \sum_{\ell \geq 0} R_\ell(s)w^\ell$ for complex w. Note that $g(s,w)$ satisfies the following formal identity

$$g(s,w) = 1 + w\mathbf{A}[g(w,\cdot)](s) = 1 + w \sum_{j \geq 0} g(s-j,w)T(s-j). \tag{40}$$

Interestingly enough, function $g(s,w)$ has a polar singularity at $w = 1/T(s)$, as proved below.

Lemma 3 *There exists a function $h(s,w)$ that is analytic for all w and s satisfying*

$$wT(s-m) \neq 1 \quad \textit{for all } m \geq 1$$

such that

$$g(s,w) = \frac{h(s,w)}{1 - wT(s)}. \tag{41}$$

Thus, $g(s,w)$ has a meromorphic continuation where $w_0 = 1/T(s)$ is a polar singularity.

Proof: We recall that $R_k(s) = \mathbf{A}^k[1](s)$. Then for $p < q$

$$|R_k(s)| \leq \frac{1}{\prod_{j \geq 1}(1-p^j)} (p^{-\Re(s)} + q^{-\Re(s)})^k.$$

Thus, if $|w| < T(\Re(s))^{-1}$ the series

$$g(s,w) = \sum_{\ell \geq 0} R_\ell(s)w^\ell = \left(\sum_{\ell \geq 0} w^\ell \mathbf{A}^\ell \right) [1](s) \tag{42}$$

converges absolutely and represents an analytic function. We can rewrite (42) as

$$g(s,w) = (\mathbf{I} - w\mathbf{A})^{-1}[1](s),$$

where $\mathbf{I}$ is the identity matrix. Then

$$(\mathbf{I} - w\mathbf{A})[g(w,\cdot)](s) = g(s,w) - w\sum_{j\geq 0} g(s-j,w)T(s-j) = 1, \tag{43}$$

which formally proves (40).

If we substitute $g(s,w)$ in (40) by

$$g(s,w) = \frac{h(s,w)}{1-wT(s)},$$

we find the following relation for $h(s,w)$

$$h(s,w) = 1 + \sum_{j\geq 1} h(s-j,w)\frac{wT(s-j)}{1-wT(s-j)}. \tag{44}$$

Recall that we establish the existence of $h(s,w)$ for $|w| < T(\Re(s))^{-1}$. We will now use (44) to show that $h(s,w)$ can be analytically continued to $|w| < T(\Re(s)-1)^{-1}$ (and even to all w such that $wT(s-m) \neq 1$)) thus leading to a meromorphic continuation, as claimed.

For this purpose we introduce another operator $\mathbf{B}$ defined as

$$\mathbf{B}[f](s) = \sum_{j\geq 1} f(s-j,w)\frac{wT(s-j)}{1-wT(s-j)}. \tag{45}$$

For convenience, set $U(s,w) = wT(s)/(1-wT(s))$. By induction it follows that

$$\begin{aligned}\mathbf{B}^k[1](s) &= \sum_{i_1\geq 1}\sum_{i_2\geq 1}\cdots\sum_{i_k\geq 1} U(s-i_1,w)U(s-i_1-i_2,w) \qquad \cdots U(s-i_1-i_2-\cdots-i_k,w)\\ &= \sum_{m_k\geq k}\sum_{m_{k-1}=k-1}^{m_k-1}\sum_{m_{k-2}=k-2}^{m_{k-1}-1} \qquad \cdots\sum_{m_1=1}^{m_2-1} U(s-m_1,w)U(s-m_2,w)\cdots U(s-m_k,w).\end{aligned}$$

Hence,

$$\begin{aligned}|\mathbf{B}^k[1](s)| &\leq \sum_{m_k\geq k}\sum_{m_{k-1}\geq k-1}\cdots \qquad \sum_{m_1\geq 1} |U(s-m_1,w)U(s-m_2,w)\cdots U(s-m_k,w)|\\ &= \sum_{m_1\geq 1}|U(s-m_1,w)|\cdot\sum_{m_2\geq 2}|U(s-m_2,w)| \qquad \cdots\sum_{m_k\geq k}|U(s-m_w,k)|.\end{aligned}$$

By using the fact that $T(s-m) = O(q^m)$, it follows directly that the series

$$S := \sum_{m\geq 1}|U(s-m,w)| = \sum_{m\geq 1}\frac{|wT(s-m)|}{|1-wT(s-m)|}$$

converges if $wT(s-m) \neq 1$ for all $m \geq 1$. Thus for any choice of w and s there are only finitely many exceptional points where $wT(s-m) = 1$.

Let now k_0 be any value such that

$$\sum_{m\ge k_0} |U(s-m,w)| \le \frac{1}{2}.$$

Then we have for all $k \ge k_0$

$$|\mathbf{B}^k[1](s)| \le S^{k_0} 2^{-(k-k_0)} = (2S)^{k_0} 2^{-k}.$$

In view of this,

$$h(s,w) = \sum_{k\ge 0} \mathbf{B}^k[1](s), \tag{46}$$

is well defined and it satisfies (44). Furthermore, $|h(s,w)| \le 2(2S)^{k_0}$. ∎

Now we are in the position to derive an asymptotic representation for $F_k(s)$.

Lemma 4 *For every real interval $[a,b]$ there exist k_0, $\eta > 0$ and $\varepsilon > 0$ such that*

$$F_k(s) = f(s)T(s)^k \left(1 + O\left(e^{-\eta k}\right)\right) \tag{47}$$

uniformly for all s such that $\Re(s) \in [a,b]$, $|\Im(s) - 2\ell\pi/\log(q/p)| \le \varepsilon$ for some integer ℓ and $k \ge k_0$, where $f(s)$ is an analytic function that satisfies $f(-r) = 0$ for $r = 1,2,\ldots$.

Furthermore, if $|\Im(s) - 2\ell\pi\log(q/p)| > \varepsilon$ for for all integers ℓ then we have

$$F_k(s) = O\left(T(\Re(s))^k\, e^{-\eta k}\right). \tag{48}$$

uniformly for $\Re(s) \in [a,b]$.

Proof: The idea of the proof is to show that the function $f(s,w)$ has a polar singularity $w = 1/T(s)$ (if the imaginary part of s is close to an integer multiple of $2\pi\log(q/p)$), and to use this property to obtain asymptotic for the coefficient $F_k(s) = [w^k]\, f(s,w)$. (In a similar way we obtain estimates for $F_k(s)$ if the imaginary part of s is not close to an integer multiple of $2\pi\log(q/p)$.) Due to the special structure of $f(s,w)$ (see (37) and (39)) we have to distinguish several cases depending on the size of $\Re(s)$.

Suppose first that $s > -r-1$ for some integer $r \ge 0$ but s is not a positive integer. Here we use the following representation (cf. (39))

$$\begin{aligned} f(s,w) &= \sum_{\ell=0}^{r} F_\ell(-r) w^\ell \frac{g(s,w)}{g(-r,w)} \\ &= \sum_{\ell=0}^{r} F_\ell(-r) w^\ell \frac{h(s,w)}{h(-r,w)} \frac{1-wT(-r)}{1-wT(s)}. \end{aligned} \tag{49}$$

By Lemma 3 the function $h(s,w)$ is analytic for $|w| < 1/T(s-1)$. By (46) it also follows that $h(s,w)$ is non-zero for real $0 < w < 1/T(s-1)$. It also follows that $h(-r,w)$ is analytic and non-zero for $0 < w < 1/T(-r-1)$. Hence, $w_0 = 1/T(s)$ is a singular point of $f(s,w)$. Since $F_k(s) = \Delta_k^*(s)/\Gamma(s)$ it follows that all values $F_k(s)$, $k \ge 0$, have the same sign. Hence, the radius of convergence of the series $\sum_{k\ge0} F_k(s) w^k$ equals $w_0 = 1/T(s)$.

In a next step we show that $f(s,w)$ has no other singularities on the radius of convergence $|w| = 1/T(s)$. Moreover, the function $f(s,w)(1-wT(s))$ continues analytically to to $|w| < 1/T(s)+\varepsilon$ for some $\varepsilon > 0$. Since all terms on the right hand side of (49), that is, $\sum_{\ell=0}^r F_\ell(-r)w^\ell$, $h(s,w)$, $h(-r,w)$, $1-wT(-r)$, and $1-wT(s)$ are analytic for $|w| < 1/T(s)+\varepsilon$, a singularity of $f(s,w)$ can only be induced by a zero of $h(-r,w)$.

Suppose first that $h(-r,w)$ has a zero w_1 with $|w_1| < 1/T(s)$. Since $h(-r,w) \ne 0$ for $0 < w < 1/T(-r-1)$ it follows that $w_1 \ne 1/T(-r)$. If we assume that $\sum_{\ell=0}^r F_\ell(-r)w_1^\ell \ne 0$, then $w = w_1$ has to be a zero of $h(s,w)$. We slightly decrease s to $s-\eta$ (for some $\eta > 0$ such that $s-\eta$ is not a positive integer) such that $h(s-\eta, w_1) \ne 0$. Then the zero $w = w_1$ of $h(-r,w)$ would induce a singularity w_1 of $f(s,w)$ with $|w_1| < 1/T(s)$ although its radius of convergence is $1/T(s-\eta) > 1/T(s) > |w_1|$. This leads to a contradiction. Hence, if $h(-r,w_1) = 0$ for some w_1 with $|w_1| < 1/T(s)$, then we also have $\sum_{\ell=0}^r F_\ell(-r)w_1^\ell = 0$. Actually, it also follows that the order of the zeroes are the same.

The above considerations also show that if $w = w_1$ is a zero of $h(-r,w)$ with $|w_1| < 1/T(-r-1)$, then w_1 is also a zero of $\sum_{\ell=0}^r F_\ell(-r)w^\ell = 0$ of the same order. Namely, if $|w_1| < 1/T(-r-1)$ then there exists a non-integral real number $s > -r-1$ with $|w_1| < 1/T(s)$ and we proceed as above.

This property shows that the only singularity of the mapping $w \mapsto f(s,w)$ is given by $w = 1/T(s)$ if $s > -r-1$ is real (but not an integer). This singularity is polar singularity of order 1. Hence, by using Cauchy's formula for a contour of integration on the circle $\gamma = \{w \in \mathbb{C} : |w| = e^\eta/T(s)\}$ and the residue theorem [4, 24] it follows that

$$\begin{aligned} F_k(s) &= \frac{1}{2\pi i}\int_\gamma f(s,w)w^{-k-1}\,dw \\ &= f(s)T(s)^k + O\left(|T(s)e^{-\eta}|^k\right), \end{aligned}$$

where

$$f(s) = \sum_{\ell=0}^r F_\ell(-r)T(s)^{-\ell}\frac{h(s,1/T(s))}{h(-r,1/T(s))}\left(1-\frac{T(-r)}{T(s)}\right)$$

These estimates are uniform for s contained in a compact interval $[a,b] \subseteq (-r-1,-r)$ (for some non-negative integer r) or in a compact interval $[a,b]$ contained in the positive real line. Furthermore, we get the same result if s is sufficiently close to the real axis. Thus, if $a \le \Re(s) \le b$ and $|\Im(s)| \le \varepsilon$ for some sufficiently small $\varepsilon > 0$ then we obtain (47). Here we have also use that fact that $f(s) \ne 0$ in this range.

Next, suppose that s is real (or sufficiently close to the real axis) and close to a negative integer $-r$, say $-r-\eta \le s \le -r+\eta$ (for some $\eta > 0$). Here we use the representation

$$\begin{aligned} \sum_{k\ge 0} F_k(s)w^k &= \sum_{\ell=0}^r F_\ell(-r)w^\ell\frac{g(s,w)}{g(-r,w)} \\ &= \sum_{\ell=0}^r F_\ell(-r)w^\ell\,\frac{h(s,w)}{h(-r,w)}\,\frac{1-wT(-r)}{1-wT(s)} \\ &= \sum_{\ell=0}^r F_\ell(-r)w^\ell\,\frac{h(s,w)-h(-r,w)}{h(-r,w)}\,\frac{1-wT(-r)}{1-wT(s)} \\ &\quad + \sum_{k=0}^r F_\ell(-r)w^\ell + \sum_{\ell=0}^r F_\ell(-r)w^{\ell+1}\frac{T(s)-T(-r)}{1-wT(s)} \end{aligned}$$

Now if we subtract the finite sum $\sum_{\ell=0}^{r} F_\ell(-r)w^\ell$, then we can safely multiply by $\Gamma(s)$ (that is singular at $s=-r$) and obtain

$$\begin{aligned}\Gamma(s)\sum_{k>r} F_k(s)w^k =& \sum_{\ell=0}^{r} F_\ell(-r)w^\ell\, \frac{\Gamma(s)(h(s,w)-h(-r,w))}{h(-r,w)}\,\frac{1-wT(-r)}{1-wT(s)} \\ &+\sum_{\ell=0}^{r} F_\ell(-r)w^{\ell+1}\frac{\Gamma(s)(T(s)-T(-r))}{1-wT(s)}.\end{aligned}$$

We again use the fact that the function $\sum_{\ell=0}^{r} F_\ell(-r)w^\ell/h(-r,w)$ is analytic for $|w|<1/T(-r-1)$ and observe that $w=1/T(s)$ is a polar singularity. By applying Cauchy's formula we obtain for $k>r$ (similarly to the above)

$$\begin{aligned}\Gamma(s)F_k(s) =& \sum_{\ell=0}^{r} F_\ell(-r)T(s)^{-\ell}\, \frac{\Gamma(s)(h(s,1/T(s))-h(-r,1/T(s)))}{h(-r,1/T(s))}\left(1-\frac{T(-r)}{T(s)}\right)T(s)^k \\ &+\sum_{\ell=0}^{r} F_\ell(-r)T(s)^{-\ell-1}\,\Gamma(s)(T(s)-T(-r))\,T(s)^k \\ &+O\left(|T(s)e^{-\eta}|^k\right).\end{aligned}$$

Thus, we actually prove (47) for $k>r$ and also observe $f(-r)=0$.

Next observe that (for integers ℓ)

$$T(s+2\pi i\ell/\log(q/p)) = e^{-2\pi i\ell\log(p)/\log(q/p)}T(s)$$

Hence, $|T(s+2\pi i\ell/\log(q/p))| = |T(s)|$ and consequently, it follows that $w=1/T(s)$ is a polar singularity of $f(s,w)$ if $|\Im(s)-2\pi i\ell/\log(q/p))|<\varepsilon$ for some integer ℓ. Thus, (47) follows also for s in this range.

Finally, if $|\Im(s)-2\ell\pi/\log(q/p)|>\varepsilon$ for some integer ℓ, then there exists $\eta>0$ such that $|T(s)|<e^{-2\eta}|T(\Re(s)|$. Hence it follows that $f(s,w)$ is regular for $|w|<e^{2\eta}/T(\Re(s))$. Hence, if we use the path of integration $\gamma=\{w\in\mathbb{C}:|w|=e^\eta/T(\Re(s))\}$ in Cauchy's formula to obtain

$$F_k(s) = O\left(T(\Re(s))^k e^{-\eta k}\right)$$

which is precisely (48). It should be clear that this estimate is uniform if $\Re(s)$ varies in a finite interval $[a,b]$. ■

Similarly we can analyze function $\overline{F}_k(s)$. Following the same footsteps, we conclude that (for $|\Im(s)-t_j|\le\varepsilon$ for some integer j)

$$\overline{F}_k(s) = \overline{f}(s)T(s)^k\left(1+O(e^{-\eta k})\right), \tag{50}$$

where $\overline{f}(-r)=0$ for $r=1,2,\ldots$.

4.2.2 Saddle Point Analysis

By the above discussion, we know that $F_k(s)$ and $\Delta_k(s)=\Gamma(s)F_k(s)$ behave asymptotically as $T(s)^k$. Thus we are in a situation similar to the analysis of the profile of random tries presented in [19]. Our asymptotic analysis will be therefore similar to that of [19].

We start with a very short outline of the proof (where we also make a simplification and we only consider the case $x = n$). By applying the inverse Mellin transform

$$\Delta_k(n) = \frac{1}{2\pi i} \int_{\rho - i\infty}^{\rho + i\infty} \Delta_k^*(s) n^{-s} \, ds \tag{51}$$

it is natural to choose $\rho = \rho_{n,k}$ as the saddle point of the function

$$T(s)^k n^{-s} = e^{k \log T(s) - s \log n}.$$

Observe that

$$\frac{k}{\log n} = \frac{p^{-\rho} + q^{-\rho}}{p^{-\rho} \log \frac{1}{p} + q^{-\rho} \log \frac{1}{q}}.$$

Note also that on the line $\Re(s) = \rho$ there will be infinitely many saddle points

$$s_j = \rho + \frac{2\pi i j}{\log \frac{p}{q}}$$

since $T(s_j) = e^{-2\pi i j (\log p)/(\log p/q)} T(\rho)$. Consequently, the behavior of $T(s)^k z^{-s}$ around $s = s_j$ is almost the same as that of $T(s)^k z^{-s}$ around $s = \rho$. This phenomenon gives a periodic leading factor in the asymptotics of $\mathbb{E}\, B_{n,k}$.

Lemma 5 *Suppose that* $(\log \frac{1}{p})^{-1} + \varepsilon \le k/\log n \le (\log \frac{1}{q})^{-1}$ *(for some* $\varepsilon > 0$*). Then*

$$\Delta_k(n) = H\left(\rho_{n,k}, \log_{p/q} p^k n\right) \frac{(p^{-\rho_{n,k}} + q^{-\rho_{n,k}})^k n^{-\rho_{n,k}}}{\sqrt{2\pi \beta(\rho_{n,k}) k}} \left(1 + O\left(\frac{1}{\sqrt{k}}\right)\right), \tag{52}$$

where

$$H(\rho, x) = \sum_{j \in \mathbb{Z}} f(\rho + it_j) \Gamma(\rho + it_j + 1) e^{-2j\pi i x} \tag{53}$$

is a non-zero periodic function with period 1 *and* $t_j = 2j\pi/\log(p/q)$.

Proof: For convenience we set $J_k(n,s) = n^{-s}\Gamma(s)F_k(s)$. By Lemma 4 we can safely replace $F_k(s)$ by $f(s)T(s)^k$ since the error term is of order $O(|T(s)|^k e^{-\eta k})$ and leads to an exponentially small contribution compared to the asymptotic leading term.

We split the integral (51) into two parts where we use the substitution $s = \rho + it$. Let us start with the range $|t| \ge \sqrt{\log n}$ and recall that by Stirling's formula $\Gamma(\rho + it) = O\left(|t|^{\rho - 1/2} e^{-\pi |t|/2}\right)$. Then

$$\begin{aligned}
\Delta_k(n) = \frac{1}{2\pi} \int\limits_{|t| \ge \sqrt{\log n}} J_k(n, \rho + it)\, dt &= O\left(n^{-\rho} T(\rho)^k \int_{\sqrt{\log n}}^{\infty} |\Gamma(\rho + it)| \, dt\right) \\
&= O\left(n^{-\rho} T(\rho)^k \int_{\sqrt{\log n}}^{\infty} t^{\rho - 1/2} e^{-\pi t/2}\right) \\
&= O\left(n^{-\rho} T(\rho)^k (\log n)^{\rho/2 - 1/4} e^{-\pi \sqrt{\log n}/2}\right) \\
&= O\left(n^{-\rho} T(\rho)^k e^{-\sqrt{\log n}}\right).
\end{aligned}$$

Next set

$$T_j = \frac{1}{2\pi}\int_{|t-t_j|\le \pi/\log(p/q)} J_k(n,\rho+it)\,dt,$$

where $t_j = \frac{2\pi j}{\log\frac{p}{q}}$. We have to study these integrals for all $|j| \le j_0 = \lfloor \sqrt{\log n}\log(p/q)/(2\pi)\rfloor$. Since there exists $c_0 > 0$ such that

$$p^{-\rho-it} + q^{-\rho-it} \le T(\rho)e^{-c_0(t-t_j)^2}$$

for $|t-t_j| \le \pi/\log(p/q)$, we obtain an upper bound of the integral (for $j \ne 0$)

$$\begin{aligned}
T_j' =& \frac{1}{2\pi}\int\limits_{k^{-2/5}\le|t-t_j|\le\pi/\log(p/q)} J_k(n,\rho+it)\,dt \\
&= O\left(|\Gamma(\rho+it_j)|n^{-\rho}T(\rho)^k\int_{k^{-2/5}}^{\infty} e^{-c_0kt^2}\,dt\right)\\
&= O\left(|\Gamma(\rho+it_j)|n^{-\rho}T(\rho)^k k^{-3/5}e^{-c_0k^{1/5}}\right).
\end{aligned}$$

For $j=0$ we can replace the factor $|\Gamma(\rho+it_j)|$ by 1.

Finally, for $|t-t_j| \le k^{-2/5}$ we use the approximation

$$\begin{aligned}
J_k(n,\rho+it) &= \Gamma(\rho+it)f(\rho+it)n^{\rho+it}T(\rho+it)^k\\
&= \Gamma(\rho+it)f(\rho+it)e^{-it_j\log(p^kn)}n^{-\rho+i(t-t_j)}T(\rho+i(t-t_j))^k\\
&= \Gamma(\rho+it_j)f(\rho+it_j)e^{-it_j\log(p^kn)}n^{-\rho}T(\rho)^k e^{-\frac12\beta(\rho)(t-t_j)^2}\\
&\times\left(1+O(|t-t_j|)+O(k|t-t_j|^3)\right).
\end{aligned}$$

A standard saddle point method then leads to

$$\begin{aligned}
T_j'' =& \frac{1}{2\pi}\int\limits_{|t-t_j|\le k^{-2/5}} J_k(n,\rho+it)\,dt\\
&= \Gamma(\rho+it_j)f(\rho+it_j)\frac{n^{-\rho}T(\rho)^k}{\sqrt{2\pi\beta(\rho)k}}e^{-it_j\log(p^kn)}\left(1+O(k^{-1/2})\right).
\end{aligned}$$

Hence we finally obtain

$$\begin{aligned}
\Delta_k(n) &= \sum_{|j|\le j_0} T_j + O\left(n^{-\rho}T(\rho)^k e^{-\sqrt{\log n}}\right)\\
&= \sum_{|j|\le j_0}\Gamma(\rho+it_j)f(\rho+it_j)e^{-it_j\log(p^kn)}\frac{n^{-\rho}T(\rho)^k}{\sqrt{2\pi\beta(\rho)k}}\left(1+O(k^{-1/2})\right)\\
&+ O\left(n^{-\rho}T(\rho)^k e^{-\sqrt{\log n}}\right)\\
&= H\left(\rho,\log_{p/q}p^kn\right)\frac{n^{-\rho}T(\rho)^k}{\sqrt{2\pi\beta(\rho_{n,k})k}}\left(1+O(k^{-1/2})\right),
\end{aligned}$$

as desired. ■

Remark 6 The above proof extends directly to an asymptotic expansion for $\Delta_k(ne^{i\vartheta})$, where $|\vartheta| \le \pi/2 - \varepsilon$ (for some $\varepsilon > 0$). In this range we have uniformly

$$\Delta_k(ne^{i\vartheta}) = \frac{T(\rho)^k}{\sqrt{2\pi\beta(\rho)k}} \sum_{|j|\le j_0} \Gamma(\rho+it_j)f(\rho+it_j)(ne^{i\vartheta})^{-\rho-it_j}p^{-ikt_j} \times \left(1 + O\left(k^{-1/2}\right)\right). \tag{54}$$

We will use this extended version for the final depoissonization procedure.

The analysis of $\overline{\Delta}_k(x)$ for the internal profile is similar but needs some additional considerations. As before, we start with

$$\overline{\Delta}_k(n) = \frac{1}{2\pi i}\int_{\rho-i\infty}^{\rho+i\infty} \overline{\Delta}_k^*(s)n^{-s}\,ds \tag{55}$$

where, we recall, $\overline{\Delta}_k^*(s) = -\overline{F}_k(s)\Gamma(s)$. In (50) we establish that $\overline{F}_k(s) = \overline{f}(s)T(s)^k\left(1+O(e^{-\eta k})\right)$, with $\overline{f}(-r) = 0$ for $r = 1, 2, \ldots$. Notice that $\overline{f}(-0) \ne 0$. In fact, we know that $\overline{F}_k(0) = 2^k$. With these preliminaries, we are ready to present our asymptotic analysis.

We have to distinguish three ranges:

Range: $\left(\log\frac{1}{p}\right)^{-1} + \varepsilon \le k/\log n \le 2\left(\log\frac{1}{p} + \log\frac{1}{q}\right)^{-1} - \varepsilon$.
In order to cover this range we have to shift the line of integration in (55) to the saddle point $\rho > 0$. By doing this we collect a contribution of 2^k from the polar singularity of $\overline{F}_k(s)\Gamma(s)$. This leads to

$$\overline{\Delta}_k(x) = 2^k + \frac{1}{2\pi i}\int_{\rho-i\infty}^{\rho+i\infty} \overline{\Delta}_k^*(s)x^{-s}\,ds.$$

The remaining integral can be handled as above by a saddle point method.

Range: $2\left(\log\frac{1}{p} + \log\frac{1}{q}\right)^{-1} + \varepsilon \le k/\log n \le \left(\log\frac{1}{q}\right)^{-1} - \varepsilon$.
Here we have $\rho < 0$ and we are precisely in the same situation as in the analysis of the external profile. Actually this range is the most significant range. Almost all nodes are concentrated around the level $k/\log n \approx 1/h$, where $h = p\log\frac{1}{p} + q\log\frac{1}{q}$ denotes the entropy of the source.

Range: $k/\log n \approx 2\left(\log\frac{1}{p} + \log\frac{1}{q}\right)^{-1}$.
Here a phase transition occurs. Technically, a polar singularity (of $\Gamma(s)$) and the saddle point $F_k(s)n^{-s}$ coalesce at $s = 0$. More precisely, we assume that $\alpha = k/\log n$ is close to $\alpha_0 := 2\left(\log\frac{1}{p} + \log\frac{1}{q}\right)^{-1}$. More precisely suppose that

$$k = \alpha_0\left(\log n + \xi\sqrt{\alpha_0\beta(-2)\log n}\right),$$

where $\xi = o((\log n)^{1/6})$. Here we move the line of integration to the saddle point

$$\Re(s) = \rho = \frac{1}{\log(p/q)}\log\frac{1-\alpha\log(1/p)}{\alpha\log(1/q)-1} = -\frac{\xi}{\sqrt{\alpha_0\beta(0)\log n}} + O\left(\xi^2/\log n\right).$$

First assume that $k > \alpha_0 \log n$, so that $\xi > 0$ and $\rho < 0$. This means that we do not pass the polar singularity, which is located at $s = 0$. Hence, as above we obtain

$$\begin{aligned}\Delta_k(ne^{i\vartheta}) &= \frac{1}{2\pi}\int_{|t|\le(\log n)^{-2/5}} \overline{J}_k(ne^{i\vartheta},\rho+it)\,dt\\ &\quad + O\left(|\Gamma(\rho+1+i(\log n)^{-2/5})|\,n^{-\rho}T(\rho)^k e^{-c_0(\log n)^{1/5}}\right)\\ &\quad + O\left(k^{-1/2}n^{-\rho}T(\rho)^k\right),\end{aligned}$$

where $\overline{J}_k(s,w) = x^{-s}\Gamma(s)\overline{F}_k(s)$. This can be again replaced by $x^{-s}\Gamma(s)\overline{f}(s)T(s)^k$. Since

$$|\Gamma(\rho+i(\log n)^{-2/5})| = O\left(\frac{1}{|\xi|(\log n)^{-1/2}+(\log n)^{-2/5}}\right) = O((\log n)^{2/5}),$$

we can neglect the first error term.

Next we replace the factor $\Gamma(s)\overline{f}(s)$ in (the approximation of) $J_k(s,w)$ by

$$\frac{\overline{f}(s)}{s}.$$

Since the sum $\Gamma(s)\overline{f}(s) - \overline{f}(0)/s$ is analytic, we have

$$\int_{|t|\le(\log n)^{-2/5}}\left(\Gamma(s)\overline{f}(s)-\frac{\overline{f}(0)}{s}\right)(ne^{i\vartheta})^{-\rho-it}T(\rho+it)^k\,dt = O\left(\frac{n^{-\rho}T(\rho)^k}{\sqrt{k}}\right)$$

and consequently

$$\begin{aligned}\Delta_k(ne^{i\vartheta}) &= \frac{\overline{f}(0)}{2\pi}\int_{|t|\le(\log n)^{-2/5}}\frac{(ne^{i\vartheta})^{-\rho-it}T(\rho+it)^k}{\rho+2+it}\,dt\\ &= \frac{\overline{f}(0)}{2\pi}n^{-\rho}e^{-i\vartheta\rho}T(\rho)^k\int_{|t|\le(\log n)^{-2/5}}\frac{e^{\vartheta t-\beta(\rho)kt^2/2+O(k|t|^3)}}{\rho+2+it}\,dt\\ &= \frac{\overline{f}(0)}{2\pi}n^{-\rho}e^{-i\vartheta\rho}T(\rho)^k\int_{-\infty}^{\infty}\frac{e^{-t^2/2}}{\xi_0+it}\left(1+O\left(\frac{|t|+|t|^3}{\sqrt{\log n}}\right)\right)dt,\end{aligned}$$

where

$$\xi_0 = (\rho+2)\sqrt{\beta(\rho)k} = -\xi + O(\xi^2(\log n)^{-1/2}).$$

Since $\xi_0 < 0$, we obtain

$$\begin{aligned}\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-t^2/2}}{\xi_0+it}\,dt &= \frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-t^2/2}\int_0^{\infty}e^{-v(\xi_0+it)}\,dv\,dt\\ &= \frac{1}{2\pi}\int_0^{\infty}e^{-v\xi_0}\int_{-\infty}^{\infty}e^{-t^2/2-itv}\,dt\,dv\\ &= \frac{1}{\sqrt{2\pi}}\int_0^{\infty}e^{-v^2/2-v\xi_0}\,dv\\ &= e^{\xi_0^2/2}\Phi(-\xi_0).\end{aligned}$$

The error term is estimated similarly:

$$\frac{1}{\sqrt{\log n}}\int_{-\infty}^{\infty}\frac{(|t|+|t|^3)e^{-t^2/2}}{|(\rho+2)\sqrt{\beta(\rho)k}+it|}\,dt = O\left(\frac{1}{\sqrt{\log n}}\int_0^{\infty}(v+v^3)e^{-v^2/2-v\xi_0}\,dv\right)$$

$$= O\left(\frac{1}{\sqrt{\log n}} e^{\xi_0^2/2}\Phi(-\xi_0)(1+|\xi_0|^3)\right).$$

Thus

$$\overline{\Delta}_k(ne^{i\vartheta}) = \overline{f}(0)(ne^{i\vartheta})^{-\rho}T(\rho)^k e^{\xi_0^2/2}\Phi(-\xi_0)\left(1+O\left(\frac{1+|\xi_0|^3}{\sqrt{\log n}}\right)\right) + O\left(k^{-1/2}n^{-\rho}T(\rho)^k\right).$$

By using the local expansions

$$\begin{aligned} n^{-\rho}T(\rho)^k &= T(0)^k e^{-\xi^2/2+O(|\xi^3|(\log n)^{-1/2})}, \\ e^{\xi_0^2/2}\Phi(-\xi_0) &= e^{\xi^2/2}\Phi(\xi)\left(1+O\left(|\xi|^3(\log n)^{-1/2}\right)\right) \end{aligned}$$

we end up with the final expansion

$$\Delta_k(ne^{i\vartheta}) = \overline{f}(0)(ne^{i\vartheta})^2 T(0)^k \Phi(\xi)\left(1+O\left(\frac{1+|\xi_0|^3}{\sqrt{\log n}}\right)\right) + O\left(k^{-1/2}T(0)^k e^{-\xi^2/2}\right),$$

that holds uniformly for $|\vartheta| \le \vartheta_0$.

4.2.3 Depoissonization

The final step in the proof is to obtain asymptotics for $\mathbb{E}\, B_{n,k}$ and $\mathbb{E}\, I_{n,k}$ from the asymptotic properties of $\Delta_k(x)$ and $\overline{\Delta}_k(x)$. This is accomplished by the analytical depoissonization [8] which requires to compute another Cauchy integral

$$\mathbb{E}\, B_{n,k} = \frac{n!}{2\pi i}\int_{|x|=n} e^x \Delta_k(x)\frac{dx}{x^{n+1}}.$$

Since $\Delta_k(x)$ behaves quite smoothly (in particular it has a subexponential growth) the depoissonization heuristics saying that $\mathbb{E}\, B_{n,k} \approx \Delta_k(n)$ applies (see [4, 24]). However, this has to be made precise. For this we need a good upper bound for $\Delta_k(ne^{i\vartheta})$ that is valid for all $|\vartheta| \le \pi$.

Lemma 6 *For every real number ρ there exists a constant $C = C(\rho)$ and an integer $k_0 = k_0(\rho)$ such that*

$$|e^x\Delta_k(x)| \le C(1-c\vartheta^2)^{-k} r^{\max\{-\rho,0\}} T(\rho)^k e^{r(1-c\vartheta^2)} \tag{56}$$

for $k \ge k_0$ and uniformly for all $r \ge 0$ and $|\vartheta| \le \pi$, where $x = re^{i\vartheta}$.

Proof: We indicate a proof for $\rho \le 0$. Obviously there exists k_0 and C such that (56) holds for $k = k_0$. Then by definition we have recursively (for $k \ge k_0$)

$$\begin{aligned} |e^x\Delta_{k+1}(x)| &= \left|\int_0^x e^{\xi}\left(\Delta_k(p\xi)+\Delta_k(q\xi)\right)\,d\xi\right| \\ &= \left|\int_0^r e^{te^{i\vartheta}}\left(\Delta_k(pte^{i\vartheta})+\Delta_k(qte^{i\vartheta})\right)\,dt\right| \\ &\le C(1-c\vartheta^2)^{-k}\,T(\rho)^k\int_0^r \left(e^{qt\cos\vartheta}(pt)^{-\rho}e^{pt(1-c\vartheta^2)} + e^{pt\cos\vartheta}(qt)^{-\rho}e^{qt(1-c\vartheta^2)}\right)\,dt \\ &\le C(1-c\vartheta^2)^{-k}T(\rho)^{k+1}\int_0^r t^{-\rho}e^{t(1-c\vartheta^2)}\,dt \end{aligned}$$

$$\leq C(1-c\vartheta^2)^{-k-1}T(\rho)^{k+1}\, r^{-\rho} e^{r(1-c\vartheta^2)}.$$

A similar proof works for $\rho \geq 0$. ■

As explained above we use Cauchy integral along $|x| = n$ to complete the asymptotic analysis:

$$\mathbb{E}\, B_{n,k} = \frac{n!}{2\pi i}\int_{|x|=n} e^x \Delta_k(x)\,\frac{dx}{x^{n+1}} = \frac{n!\, n^{-n}}{2\pi}\int_{|\vartheta|\leq\pi} e^{ne^{i\vartheta}}\Delta_k(ne^{i\vartheta})e^{-in\vartheta}\, d\vartheta.$$

Fix $0 < \vartheta_0 < \pi/2$. Then Lemma 6 implies

$$\begin{aligned}\left|\frac{n!\, n^{-n}}{2\pi}\int\limits_{\vartheta_0\leq|\vartheta|\leq\pi} e^{ne^{i\vartheta}}\Delta_k(ne^{i\vartheta})e^{-in\vartheta}\, d\vartheta\right| &\leq \Delta_k(n)\frac{n!\, n^{-n}e^n}{2\pi}\int\limits_{\vartheta_0\leq|\vartheta|\leq\pi} e^{-cn\vartheta^2}\, d\vartheta \\ &= O\left(\Delta_k(n)e^{-c\vartheta_0^2 2n}\right).\end{aligned}$$

For the remaining part of the integral we use (54) and obtain

$$\begin{aligned}&\frac{n!\, n^{-n}}{2\pi}\int_{|\vartheta|\leq\vartheta_0} e^{ne^{i\vartheta}}\Delta_k(ne^{i\vartheta})e^{-in\vartheta}\, d\vartheta \\ &\quad= \frac{n^{-\rho}T(\rho)^k}{\sqrt{2\pi\beta(\rho)k}}\sum_{|j|\leq j_0}\Gamma(\rho+it_j)f(\rho+it_j) \\ &\quad\times\frac{n!\, n^{-n}}{2\pi}\int_{|\vartheta|\leq\vartheta_0} e^{ne^{i\vartheta}-in\vartheta}e^{i\vartheta(\rho+it_j)}\, d\vartheta\cdot\left(1+O(k^{-1/2})\right) \\ &\quad= \frac{n^{-\rho}T(\rho)^k}{\sqrt{2\pi\beta(\rho)k}}\sum_{|j|\leq j_0}\Gamma(\rho+it_j)f(\rho+it_j) \\ &\quad\times\frac{n!\, n^{-n}e^n}{2\pi}\int_{|\vartheta|\leq\vartheta_0} e^{-\frac{1}{2}n\vartheta^2}\left(1+O(n|\vartheta|^3)+O(|t_j\vartheta|)\right)\, d\vartheta\cdot\left(1+O(k^{-1/2})\right) \\ &\quad= \frac{n^{-\rho}T(\rho)^k}{\sqrt{2\pi\beta(\rho)k}}\sum_{|j|\leq j_0}\Gamma(\rho+it_j)f(\rho+it_j)\left(1+O(|t_j|n^{-1/2})+O(k^{-1/2})\right) \\ &\quad= \Delta_k(n)\left(1+O(k^{-1/2})\right).\end{aligned}$$

This completes the proof of Theorem 4. The last part of the proof of Theorem 5 follows the same footsteps and is omitted.

References

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

[2] L. Devroye and H.-K. Hwang, Width and mode of the profile for some random trees of logarithmic height, *Ann. Appl. Probab.*, 16, 886–918, 2006.

[3] P. Flajolet, X. Gourdon, and P. Dumas, Mellin transforms and asymptotics: harmonic sums, Theoret. Comput. Sci. 144, 3–58, 1995.

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

[5] G. Gasper and M. Rahman, *Basic Hypergeometric Series*, Encyclopedia of Mathematics and its Applications, Vol. 96, Second Edition, (Cambridge University Press, Cambridge, 2004).

[6] P. Jacquet and W. Szpankowski, Analysis of digital tries with Markovian dependency, IEEE Trans. Information Theory 37, 1470–1475, 1991.

[7] P. Jacquet, and W. Szpankowski, Asymptotic Behavior of the Lempel-Ziv Parsing Scheme and Digital Search Trees, *Theoretical Computer Science*, 144, 161–197, 1995.

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

[9] P. Jacquet and M. Régnier, Trie partitioning process: limiting distributions, Lecture Notes in Comput. Sci. 214, 196–210, Springer, Berlin, 1986.

[10] 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.

[11] C. Knessl and W. Szpankowski, On the Average Profile of Symmetric Digital Search Trees, *Analytic Combinatorics*, 4, article # 6, 2009.

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

[13] Lothaire, M. *Applied Combinatorics on Words.* Cambridge University Press, New York, 2005

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

[15] G. Louchard and W. Szpankowski, Average Profile and Limiting Distribution for a Phrase Size in the Lempel-Ziv Parsing Algorithm, *IEEE Trans. Information Theory*, 41, 478–488, 1995.

[16] H. Mahmoud, Evolution of random search trees, John Wiley & Sons Inc., New York, 1992.

[17] M. Naor and U. Wieder, Novel Architectures for P2P Applications: The Continuous-discrete Approach, *Proceedings of the 15th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA 2003)*, 50–59, 2003.

[18] G. Park Profile of Tries, Ph.D. Thesis, Purdue University, 2006.

[19] G. Park, H.K. Hwang, P. Nicodeme, and W. Szpankowski, Profile of Tries, *SIAM J. Computing*, to appear; also *Proc. LATIN'08*, LNCS 4957, 1-11, 2008.

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

[21] Helmut Prodinger, Digital Search Trees and Basic Hypergeometric Functions, *Bulletin of the EATCS*, 56, 1995.

[22] R. Sedgewick, *Algorithms in C: Fundamental Algorithms, Data Structures, Sorting, Searching*, Addison-Weseley, 1997.

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

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