# (Un)Expected Behavior of Digital Search Tree Profile *


Michael Drmota† Wojciech Szpankowski‡



## Abstract

A digital search tree (DST) – one of the most fundamental data structure on words – is a digital tree in which keys (strings, words) are stored directly in (internal) nodes. Such trees find myriad of applications from the popular Lempel-Ziv'78 data compression scheme to distributed hash tables. 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., height, fill-up) can be expressed in terms of the profile. However, from the inception of DST, the analysis of the profile has been elusive and it has become a prominent open problem in the area of analysis of algorithms. We make here the first, but decisive step, towards solving this problem. We present a precise analysis of the average profile when stored strings are generated by a biased memoryless source. The main technical difficulty of analyzing the profile lies in solving a sophisticated recurrence equation. We present such a solution for the Poissonized version of the problem (i.e., when the number of stored strings is generated by a Poisson distribution) in the Mellin transform domain. To accomplish it, we introduce a novel functional operator that allows us to express the solution in an explicit form, and then using analytic algorithmics tools to extract asymptotic behavior of the profile. 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 it oscillates in this range. Our results are derived by methods of analytic algorithmics such as generating functions, Mellin transform, Poissonization and de-Poissonization, the saddle-point method, singularity analysis and uniform asymptotic analysis.



*The work of this author was supported in part by the Austrian Science Foundation FWF Grant No. S9604, and by the NSF Grants CCF-0513636, DMS-0503742, CCF -0830140, and DMS-0800568, NIH Grant R01 GM068959-01, NSA Grant H98230-08-1-0092, and the AFOSR Grant FA8655-08-1-3018. This work was completed during a visit at Hewlett-Packard Laboratories, Palo Alto, CA.



†Inst. Discrete Mathematics and Geometry, TU Wien, A-1040 Wien, Austria, michael.drmota@tuwien.ac.at.

‡Department of Computer Science, Purdue University, West Lafayette, IN 47907-2066 U.S.A., spa@cs.purdue.edu



*Index Terms*: Digital search trees, trees profile, analytic combinatorics, analysis of algorithms, generating functions, Mellin transform.


## 1 Introduction

*Digital trees* are fundamental data structures on words [6, 15, 17]. Among them *tries* and *digital search trees* stand out due to a myriad of applications ranging from data compression to distributed hash tables [9, 17]. In a digital search trees, the subject of this paper, strings are directly stored in nodes. More precisely, the root contains the first 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. The search for an available node follows the prefix structure of a new string [15]. In this paper, we are concerned with probabilistic properties of the *profile* defined as the sequence of numbers each counting the number of nodes with the same distance from the root. Throughout the paper, we write $X_{n,k}$ for the number of nodes at level $k$ when $n$ strings are stored (cf. Figure 1). 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 $q := 1-p > p$. 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.

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* [4]), to telecommunication (e.g., conflict resolution algorithms [17]), to partial matching of multidimensional data [15], to distributed hash tables [9]. Second, the profile is a fine shape measure closely connected to many other cost measures, as discussed in some depth below. Third, not only the analytic problems are mathematically chal-

*In particular, $X_{n,k}$ represents the number of phrases of length $k$ in the Lempel-Ziv'78 built over $n$ phrases.

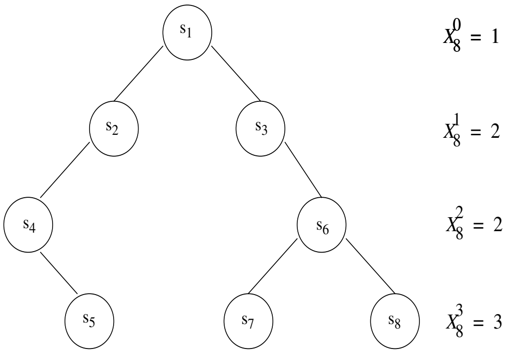


Figure 1: A digital search tree built on eight strings $s_1, \ldots, s_8$ (i.e., $s_1 = 0\ldots$, $s_2 = 1\ldots$, $s_3 = 01\ldots$, $s_4 = 11\ldots$, etc.) and its profile.

lenging, but the diverse new phenomena they exhibit are highly interesting and unusual. Fourth, our findings imply several new results on other shape parameters.

As we mentioned above, almost all DST parameters can be expressed in terms of the profile $X_{n,k}$: (i) *height*: the length of the longest path from the root becomes $H_n = \max\{j : X_{n,j} > 0\}$; (ii) *fill-up (or saturation) level*: the largest full level, or $F_n = \max\{j : X_{n,j} = 2^j\}$; (iii) *depth*: the distance from the root to a randomly selected node; its distribution is given by the expected profile divided by $n$, [8]; (iv) *total path length*: the sum of distances between nodes and the root, or equivalently, $L_n = \sum_j jX_{n,j}$.

The major difference between most previous study and the current paper is that we are dealing with asymptotics of a bivariate recurrence, in contrast to univariate recurrences addressed in the literature. The main novel mathematical result concerns an explicit and asymptotic solution of the following recurrence, never studied in the past,

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

with suitable initial conditions. The Poisson generating function $\Delta_k(z) := e^{-z} \sum_n x_{n,k} z^n/n!$, satisfies the following functional equation

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

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

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

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 which leads to infinite number of saddle points, a rather unexpected situation (cf. also [12]). The final step is to invert asymptotics of the Poisson function $\Delta_k(z)$ through the so called *analytic depoissonization* to recover asymptotically $x_{n,k}$. The reader is referred to [3, 17] for a detailed discussion of the above mentioned tools that belong to analytic algorithmics.

Digital trees have been intensively studied for the last thirty years [6, 17], 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 [6, 7, 8, 16, 14] 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 [10, 11, 12], while the profile of the digital search trees for *unbiased source* (i.e., $p = q = 1/2$) has been recently obtained in [5] (cf. Section 6.3 of Knuth [6] 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 analyze precisely the expected profile of the biased digital search tress for $k \le (\log \frac{1}{q})^{-1} \log n$ and reveal 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. Furthermore, the expected profile is oscillating in a range where profile grows polynomially. These oscillations are due to infinite number of saddle points. Knowing the expected profile for all values of depth $k$, 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. In particular, our result shows that (biased) digital search trees behave almost the same as (biased) tries.

The paper is organized as follows. We first present our main results. Then we describe a streamlined analysis with details delayed till the last two sections.

## 2 Main Results

Let $X_{n,k}$ denote the (random) number of nodes at level $k$ in a digital search tree, when $n$ strings are generated by a memoryless source with parameters $p < q = 1 - p$. It is easy to see that the probability generating function

$\mathbb{E}\, u^{X_{n,k}}$ satisfies the following recurrence relation

$$\mathbb{E}\, u^{X_{n+1,k+1}} = \sum_{\ell=0}^{n} \binom{n}{\ell} p^\ell q^{n-\ell} \mathbb{E}\, u^{X_{n,\ell}} \mathbb{E}\, u^{X_{n,n-\ell}}, \tag{2.1}$$

while the corresponding exponential generating function

$$G_k(z,u) = \sum_{n\geq 0} \mathbb{E}\, u^{X_{n,k}} \frac{z^n}{n!}$$

satisfies the following functional recurrence

$$\frac{\partial}{\partial z} G_{k+1}(z,u) = G_k(pz,u) G_k(qz,u) \tag{2.2}$$

with initial conditions $G_0(z,u) = 1 + u(e^z - 1)$ and $G_k(0,u) = 1$. We are interested in the expected profile $\mu_{n,k} = \mathbb{E}\, X_{n,k}$. By taking derivatives with respect to $u$ and setting $u = 1$ we obtain for the exponential generating function

$$E_k(z) = \sum_{n\geq 0} \mu_{n,k} \frac{z^n}{n!} = \sum_{n\geq 0} \mathbb{E}\, X_{n,k} \frac{z^n}{n!}.$$

the following functional recurrence

$$E'_{k+1}(z) = e^{qz} E_k(pz) + e^{pz} E_k(qz) \tag{2.3}$$

with initial condition $E_0(z) = e^z - 1$ and $E_k(0) = 0$.

It is known that this kind of recurrence is rather difficult to solve. In what follows we present a method to solve equations of that kind based on a three step procedure. We first apply the Poisson transform, then the Mellin transform and finally another power series representation. Each of these steps has to be properly inverted with help of analytic techniques; we will describe the road map of this procedure in the next section.

In order to state our main result we need the following notations. For a real number $\alpha$ with $(\log \frac{1}{p})^{-1} < \alpha < (\log \frac{1}{q})^{-1}$, let $\rho = \rho(\alpha)$ be defined by the equation

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

Furthermore, we set

$$\beta(\rho) = \frac{p^{-\rho} q^{-\rho} \log(p/q)^2}{(p^{-\rho} + q^{-\rho})^2}, \qquad \alpha_0 = \frac{2}{\log \frac{1}{p} + \log \frac{1}{q}}.$$

In this paper, we prove the following main findings.

THEOREM 2.1. *Let $\mathbb{E}\, X_{n,k}$ denote the expected profile in (asymmetric) digital search trees with underlying probabilities $0 < p < q = 1 - p$. 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:*

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

$$\begin{aligned}\mathbb{E}\, X_{n,k} = 2^k \quad - \quad & G\left(\rho_{n,k}, \log_{p/q} p^k n\right) \\ & \cdot \frac{(p^{-\rho_{n,k}} + q^{-\rho_{n,k}})^k n^{-\rho_{n,k}}}{\sqrt{2\pi \beta(\rho_{n,k}) k}} \\ & \cdot \left(1 + O\left(\frac{1}{\log n}\right)\right),\end{aligned}$$

*where $G(\rho, x)$ is a non-zero periodic function with period 1 and small amplitude (cf. Figure 2).*

*(ii) 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*

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

*where $\Phi(x)$ denotes the normal distribution function.*

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

$$\begin{aligned}\mathbb{E}\, X_{n,k} \quad = \quad & G\left(\rho_{n,k}, \log_{p/q} p^k n\right) \frac{(p^{-\rho_{n,k}} + q^{-\rho_{n,k}})^k n^{-\rho_{n,k}}}{\sqrt{2\pi \beta(\rho_{n,k}) k}} \\ & \cdot \left(1 + O\left(\frac{1}{\log n}\right)\right)\end{aligned}$$

*with $G(\rho, x)$ as above in (i).*

Note 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, the behavior of $\mathbb{E}\, X_{n,k}$ is governed by a power of $n$ depending on the ratio $\alpha = k/\log n$: Up to level $k = \alpha_0 \log n$, the digital search tree is almost full (i.e., has almost $2^k$ nodes) with some fluctuation contributing to the second order term. A phase transition occurs around $\alpha = \alpha_0 + O(1/\sqrt{\log})$, and for $\alpha > \alpha_0$ the profile grows polynomially oscillating around $n^{\alpha \log(p^{-\rho}+q^{-\rho})-\rho}$.

More interestingly, the average profile allows us to deriving several new and old results in a uniform manner. Let us start with the typical depth $D_n$ which is given by $P(D_n = k) = \mu_{n,k}/n$. Using Theorem 2.1(ii) around $k = 1/h \log n + cx\sqrt{\log n}$ we obtain the following surprising variant of the Local Limit Theorem for the depth.

For $k = (1/h) \log n + x\sqrt{c \log n}$ we have

$$P(D_n = k) \quad = \quad G_1\left(-1; \log_{p/q} p^k n\right) \frac{e^{-x^2/2}}{\sqrt{2\pi c \log n}}$$

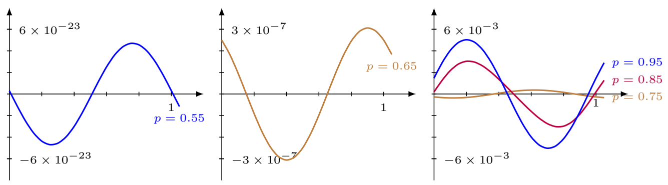


Figure 2: *The fluctuating part of the periodic function $G_1(-1;x)$ for $p = 0.55, 0.65, \ldots, 0.95$ and for $x$ in the unit interval; its amplitude tends to zero when $p \to 0.5^+$.*

$$\cdot \left(1 + O\left(\frac{1+|x|^3}{\sqrt{\log n}}\right)\right),$$

where

$$c = \frac{\beta(-1)}{h} = \frac{pq\log(p/q)^2}{p\log\frac{1}{p} + q\log\frac{1}{q}}.$$

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

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

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 on the second moment $\mathbb{E}\, X_{n,k}^2$, compare with [2].

## 3 Road Map of the Proof

As already mentioned, the proof of Theorem 2.1 consists of three steps:

(i) The **starting point** is the recurrence

(3.4)
$$\mathbb{E}\, X_{n+1,k+1} = \sum_{\ell=0}^{n} \binom{n}{\ell} p^\ell q^{n-\ell} (\mathbb{E}\, X_{\ell,k} + \mathbb{E}\, X_{n-\ell,k}) \qquad (n,k \ge 0)$$

for the expected values $\mu_{n,k} = \mathbb{E}\, X_{n,k}$. We recall the initial conditions

$$\mathbb{E}\, X_{n,0} = \begin{cases} 0 & \text{for } n = 0, \\ 1 & \text{for } n \ge 1. \end{cases}$$

(ii) The **first step** is to consider the *Poisson transform*

$$\Delta_k(z) = \sum_{n\ge 0} \mathbb{E}\, X_{n,k}\, e^{-z} \frac{z^n}{n!} = E_k(z) e^{-z} \qquad (k \ge 0)$$

that can be considered as the expected number of nodes at level $k$ if the number $n$ of total nodes follows a Poisson distribution with parameter $z$. It is clear that the above recurrence translates to

(3.5)
$$\Delta_{k+1}(z) + \Delta'_{k+1}(z) = \Delta_k(pz) + \Delta_k(qz) \qquad (k \ge 0).$$

with initial conditions $\Delta_0(z) = 1 - e^{-z}$. It is easy to prove by induction that $\Delta_k(z)$ can be represented as a finite linear combination of function of the form $e^{-p^{\ell_1} q^{\ell_2} z}$ with $\ell_1, \ell_2 \ge 0$ and $\ell_1 + \ell_2 \le k$. We will use this observation in the sequel.

(iii) The **second step** is to use the Mellin transform

$$M_k(s) = \mathcal{M}(\Delta_k(z)) = \int_0^\infty \Delta_k(z) z^{s-1}\, dz.$$

Since $\mathbb{E}\, X_{n,k} \le 2^k$ it is clear that $M_k(s)$ can only exist for $s$ with $\Re(s) < 0$. Furthermore, $X_{n,k} = 0$ for $n \le k$. Thus, $E_k(z) = O(z^{k+1})$ for $z \to 0$ which ensures that $M_k(s)$ exists for $s$ with $\Re(x) > -k-1$. Consequently $M_k(s)$ exists for $-k-1 < \Re(s) < 0$. Since $\Delta_k(z)$ can be represented as a finite linear combination of function of the form $e^{-p^{\ell_1} q^{\ell_2} z}$ with $\ell_1, \ell_2 \ge 0$ and $\ell_1 + \ell_2 \le k$ we can rewrite $M_k(s)$ as

$$M_k(s) = -\Gamma(s) F_k(s),$$

where $\Gamma(s)$ is the Euler gamma function. Observe that $F_k(s)$ is now a finite linear combination of functions of the form $p^{-\ell_1 s} q^{-\ell_2 s}$ with $\ell_1, \ell_2 \ge 0$ and $\ell_1 + \ell_2 \le k$. Thus, $F_k(s)$ can be considered as an entire function. Further, the relation (3.5) now translates to

(3.6)
$$F_{k+1}(s) - F_{k+1}(s-1) = (p^{-s} + q^{-s}) F_k(s) \qquad (k \ge 0)$$

with initial condition $F_0(s) = 1$. Note that the relation (3.6) holds not only for $-k-1 < \Re(s) < 0$ where the Mellin transform exists. Since $F_k(s)$ analytically continues to an entire function (3.6) holds for all $s$.

(iv) The **third step** is to consider the power series

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

It turns out that $f(x,s)$ can we rewritten (see Lemma 4.1) as

$$f(x,s) = \frac{g(x,s)}{g(x,-1)},$$

where $g(x,s)$ satisfies the following relation

(3.7)
$$g(x,s) = 1 + x \sum_{j\ge 0} g(x, s-j)(p^{-s+j} + q^{-s+j}).$$

Granted the above, an asymptotic analysis follows. We start with a singularity analysis of $g(x,s)$, in particular we will show (see Lemma 4.3) that $g(x,s)$ has (usually) a polar singularity at $x = 1/(p^{-s} + q^{-s})$. Thus it will be possible to get proper asymptotics for $F_k(s)$. In fact we get $F_k(s) \sim f(s)(p^{-s} + q^{-s})^k$ (for $s$ in the interesting range). This resembles an exact expression for tries of the form $(p^{-s}+q^{-s})^k$ as discussed in [12]. This is the reason why the *overall behaviour* of the profile of biased tries and biased digital search trees is almost of the same form. Only the periodic functions are slightly different.

Thus, the final two steps (inverting the Mellin transform and depoissonization) are almost identical to the methods presented in [12]. First one has to invert the Mellin transform with help of the analytic formula (through an application of the saddle point method)

$$\Delta_k(z) = -\frac{1}{2\pi i}\int_{\rho-i\infty}^{\rho+i\infty} \Gamma(s)F_k(s)z^{-s}\,ds, \tag{3.8}$$

where $-k-1<\rho<0$ and where we assume that $z$ in cone around the real axis. Finally, one has to apply analytic depoissonization to $\Delta_k(z)$ which gives $\mathbb{E}\,X_{n,k} \sim \Delta_k(n)$ Thus, for our final result we have to set $z=n$.

## 4 Singularity Analysis

Before we study the generating function $f(x,s) = \sum_{k\ge 0} F_k(s)x^k$, we will collect some basic properties of $F_k(s)$. We recall that $F_k(s)$ can be considered as entire functions.

Let $\mathbf{A}$ be an functional operator that is defined by

$$\mathbf{A}[f](s) = \sum_{j\ge 0} f(s-j)T(s-j), \tag{4.9}$$

where

$$T(s) = p^{-s}+q^{-s}. \tag{4.10}$$

In the next lemma, proved in Appendix A, we find an explicit representation of $F_k(z)$ through the operator $\mathbf{A}$.

Lemma 4.1. *The functions $F_k(s)$ are recursively given by*

$$F_{k+1}(s) = \mathbf{A}[F_k](s) - \mathbf{A}[F_k](-1) \qquad (k\ge 0). \tag{4.11}$$

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

$$\sum_{k\ge 0} F_k(s)x^k = \frac{\sum_{\ell\ge 0} R_\ell(s)x^\ell}{\sum_{\ell\ge 0} R_\ell(-1)x^\ell} \tag{4.12}$$

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

Remark 4.1. Note that if $f$ is a finite linear combination of functions of the form $p^{-\ell_1 s}q^{-\ell_2 s}$ then $\mathbf{A}[f](s)=0$ implies $f(s)=0$. This follows from the observation that

$$\mathbf{A}[p^{-\ell_1 s}q^{-\ell_2 s}] = \frac{p^{-(\ell_1+1)s}q^{-\ell_2 s}}{1-p^{\ell_1+1}q^{\ell_2}} + \frac{p^{-\ell_1 s}q^{-(\ell_2+1)s}}{1-p^{\ell_1}q^{\ell_2+1}}. \tag{4.13}$$

Thus, the largest non-zero term of $f$ (for $s\to\infty$) will be mapped into two non-zero term that contains the largest one of $\mathbf{A}[f]$.

Remark 4.2. Observe further that the proof of (4.11) (and consequently that of (4.12)) makes use of the fact that $F_k(-1)=0$ for $k\ge 1$. However, we also have $F_k(-r)=0$ for $k\ge r$. In particular, if we set $s=-r$ in (4.12) we get

$$\sum_{k=0}^{r-1} F_k(-r)x^r = \frac{\sum_{\ell\ge 0} R_k(-r)x^k}{\sum_{\ell\ge 0} R_k(-1)x^k}$$

and consequently

$$\sum_{k\ge 0} F_k(s)x^k = \frac{\sum_{\ell\ge 0} R_k(s)x^k}{\sum_{\ell\ge 0} R_k(-r)x^k}\sum_{k=0}^{r-1} F_k(-r)x^r. \tag{4.14}$$

Furthermore, since $F_k(0)=2^k$ we similarly we find

$$\sum_{k\ge 0} F_k(s)x^k = \frac{\sum_{\ell\ge 0} R_k(s)x^k}{\sum_{\ell\ge 0} R_k(0)x^k}\frac{1}{1-2x}. \tag{4.15}$$

Our next goal is to study the function $g(x,s) = \sum_{\ell\ge 0} R_\ell(s)x^\ell$, where we now consider $x$ as a complex variable, too. Note that $g(x,s)$ satisfies the (at the moment formal) identity

$$g(x,s) = 1 + x\mathbf{A}[g(x,\cdot)](s) = 1 + \sum_{j\ge 1} g(x,s-j)T(s-j). \tag{4.16}$$

In the next lemma, proved in Appendix B, we establish a crucial property of $g(x,s)$.

Lemma 4.2. *There exists a function $h(x,s)$ that is analytic for all $x$ and $s$ for which*

$$xT(s-m)\ne 1 \quad \textit{for all } m\ge 1.$$

*such that*

$$g(x,s) = \frac{h(x,s)}{1-xT(s)}. \tag{4.17}$$

*Thus, $g(x,s)$ has a meromorphic continuation where $x_0 = 1/T(s)$ is a polar singularity.*

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

Lemma 4.3. *For every real interval $[a,b]$ there exist $k_0$, $\eta>0$ and $\varepsilon>0$ such that*

$$F_k(s) = f(s)T(s)^k\left(1+O\left(e^{-\eta k}\right)\right) \tag{4.18}$$

*uniformly for all $s$ with $\Re(s)\in[a,b]$, $|\Im(s) - 2\ell\pi\log(q/p)|\le\varepsilon$ for some integer $\ell$ 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 $\ell$ then we have*

$$F_k(s) = O\left(T(\sigma)^k\, e^{-\eta k})\right). \tag{4.19}$$

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

*Proof.* Suppose first that $s$ is a real number with $-r-1<s<-r$ for some integer $r\ge 0$. Here we use the representation

$$\begin{aligned} f(x,s) &= \sum_{k=0}^{r} F_k(-r-1)x^k \frac{g(s,x)}{g(-r-1,x)} \\ &= \sum_{k=0}^{r} F_k(-r-1)x^k \frac{h(s,x)}{h(-r-1,x)} \frac{1-xT(-r-1)}{1-xT(s)}. \end{aligned}$$

By Lemma 4.2 there exist $\eta>0$ such that $h(s,x)$ is analytic for $|x|\le e^{\eta}/T(s)$. Since $T(-r-1)<T(s)$ it also follows that $h(-r-1,x)$ is analytic in that region. Furthermore, since $h(-r-1,x)$ is non-zero for positive real $x<1/T(-r-2)$ (compare with (5.26)) we obtain that the radius of convergence of the series $\sum_{k\ge0}F_k(s)x^k$ equals $x_0=1/T(s)$.

With help of this observation we can also deduce that the function $f(x,s)$ has no other singularities on the circle $|x|=1/T(s)$. Suppose that $h(-r-1,x)$ has a zero $x_1$ with $|x_1|<1/T(s)$. If $\sum_{k=0}^{r}F_k(-r-1)x_1^k\ne0$ then $x_1$ has to be a zero of $h(x,s)$, too: $h(x_1,s)=0$. However, if we slightly decrease $s$, then certainly $h(x_1,s-\eta)\ne0$. In this case the function $f(x,s)$ would be singular for $x=x_1$ although its radius of convergence is $1/T(s-\eta)>1/T(s)>|x_1|$. This is, of course, a contradiction and, thus, $\sum_{k=0}^{r}F_k(-r-1)x_1^k=0$, too. Actually, it also follows that the order of the zeroes are the same. Furthermore, by a slight variation of the above argument, we also deduce that $f(x,s)$ has no singularities on the circle $|x|=1/T(s)$ other than $x_0=1/T(s)$, as proposed.

Hence, by using a contour integration on the circle $|x|=e^{\eta}/T(s)$ and the residue theorem [3, 17] it follows that

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

where

$$\begin{aligned} f(s) &= \sum_{k=0}^{r} F_k(-r-1)T(s)^{-k}\frac{h(s,1/T(s))}{h(-r-1,1/T(s))} \\ &\quad\cdot\left(1-\frac{T(-r-1)}{T(s)}\right) \end{aligned}$$

These estimates are uniform for $s\in[a,b]$, where $-r-1<a<b<r$. 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 (4.18), too.

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

$$\begin{aligned} \sum_{k\ge0}F_k(s)x^k &= \sum_{k=0}^{r-1}F_k(-r)x^k\frac{g(s,x)}{g(-r,x)} \\ &= \sum_{k=0}^{r-1}F_k(-r)x^k\frac{h(s,x)}{h(-r,x)}\frac{1-xT(-r)}{1-xT(s)} \\ &= \sum_{k=0}^{r-1}F_k(-r)x^k\frac{h(s,x)-h(-r,x)}{h(-r,x)}\frac{1-xT(-r)}{1-xT(s)} \\ &\quad+\sum_{k=0}^{r-1}F_k(-r)x^k+\sum_{k=0}^{r-1}F_k(-r)x^{k+1}\frac{T(s)-T(-r)}{1-xT(s)} \end{aligned}$$

Now if we substract the finite sum $\sum_{k=0}^{r-1}F_k(-r)$, then we can safely multiply by $\Gamma(s)$ (that is singular at $s=-r$) and obtain a function of the form

$$\begin{aligned} &\sum_{k=0}^{r-1}F_k(-r)x^k\frac{\Gamma(s)(h(s,x)-h(-r,x))}{h(-r,x)}\frac{1-xT(-r)}{1-xT(s)} \\ &+\sum_{k=0}^{r-1}F_k(-r)x^{k+1}\frac{\Gamma(s)(T(s)-T(-r))}{1-xT(s)} \end{aligned}$$

which we can now handle in the same way as above. Thus, we actually prove (4.18) for $k\ge r$ with $f(-r)=0$.

If $s$ is close to 0 then we argue similarly. Here we can use the representation (4.15) to obtain

$$\sum_{k\ge0}F_k(s)x^k = \frac{h(s,x)}{h(0,x)}\frac{1}{1-xT(s)} \tag{4.20}$$

and (4.18) follows, too.

Finally, if $\Re(s)$ is positive (and $\Im(s)$ sufficiently close to $2\ell\pi/\log(q/p)$ for some integer $\ell$), then we can also use (4.20) and obtain the proposed result. (Note that $h(0,x)$ is analytic for $|x|<1/T(-1)<e^{\eta}|1/T(s)|$.)

Next suppose that $s=\sigma+it$, where $t$ is not necessarily small. Then

$$T(s) = e^{it\log p}\left(p^{-\sigma}+q^{-\sigma}e^{it\log(q/p)}\right).$$

Consequently $|T(s)|=T(\rho)$ if and only if $t=2k\pi/\log(q/p)$ for some integer $k$. Hence, if $|t-2k\pi/\log(q/p)|\le\varepsilon$ for some integer $k$ we can do the same contour integration as above and get again (4.18).

Finally, if $|t - 2\ell\pi/\log(q/p)| > \varepsilon$ for some integer $\ell$, then we estimate $F_k(s)$ trivially by

$$|F_k(s)| \le \rho^{-k} \cdot \max_{|x|=\rho} |g(x,s)|,$$

where $R$ is chosen in a way that $g(x,s)$ is analytic for $|x| \le R$. Since there is $\eta > 0$ with

$$|T(s-m)| = |p^{-\sigma+m} + e^{it\log(q/p}q^{-\sigma+m}| \le e^{-2\eta}T(\sigma-m)$$

it follows that $h(x,s)$ exists for $|x| \le e^{\eta}/T(\sigma)$. Hence, we can actually set $R = e^{\eta}/T(\sigma)$ and obtain (4.19). In order to complete the proof note that $M_k(s) = -\Gamma(s)F_k(s)$ exists for $-k-1 < \Re(s) < 0$ and that $F_k(-r) = 0$ for $r = 1, 2, \ldots$ and $k \ge r$. Thus, $f(-r) = 0$, too.

## 5 Saddle Point Method

By the above discussion, we know that $F_k(s)$ behaves asymptotically as $T(s)^k$. Therefore, the saddle point analysis as well as the depoissonization, is similar to those given in [12]. Thus, we will only give a very short outline of the proof. We also make a simplification that we only consider the case $z = n$.

First, for inverting the Mellin transform with (3.8) at $z = n$ it is natural to choose $\rho = \rho_{n,k}$ as the saddle point of the function

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

that is given by the relation

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

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

$$s_k = \rho + \frac{2\pi i k}{\log\frac{p}{q}}$$

since $T(s_k) = e^{-2\pi i k(\log p)/(\log p/q)}T(\rho)$ and consequently the behavior of $T(s)^k z^{-s}$ around $s = s_k$ 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 $\mu_{n,k} = \mathbb{E}\, X_{n,k}$.

We now set $\alpha = \alpha_{n,k} = k/\log n$. Recall that our goal is to derive asymptotics of $\mathbb{E}\, X_{n,k}$ for

$$\frac{1}{\log\frac{1}{p}} < \alpha < \frac{1}{\log\frac{1}{q}}.$$

In particular we distinguish between several ranges:

**Range 1**: $\frac{1}{\log\frac{1}{p}} < \alpha < \frac{2}{\log\frac{1}{p}+\log\frac{1}{q}}$.
In order to cover this range we have to shift the line of integration in (3.8) to the saddle point $\rho > 0$. By doing this we get a contribution of $2^k$ from the polar singularity of $F_k(s)\Gamma(s)$ (note that $F_k(0) = 2^k$) which is in fact the leading term. The remaining part comes from a saddle point method that evaluates (3.8) asymptotically. Note that the digital search tree is almost a complete tree in this range since the term $2^k$ dominates.

**Range 2**: $\alpha = \frac{2}{\log\frac{1}{p}+\log\frac{1}{q}}$.
Here a phase transition occurs. Technically, a polar singularity (of $\Gamma(s)$) and the saddle point $F_k(s)n^{-s}$ coalesce at $s = 0$.

**Range 3**: $\frac{2}{\log\frac{1}{p}+\log\frac{1}{q}} < \alpha < \frac{1}{\log\frac{1}{q}}$.
This is the most significant range. Almost all nodes are concentrated around the level $\alpha = 1/h$, where $h = p\log\frac{1}{p} + q\log\frac{1}{q}$ denotes the entropy of the source. This range corresponds to saddle points $\rho < 0$. Here we have to be a little bit more careful due to the polar singularities of $\Gamma(s)$ for negative integers $s$. But Lemma 4.3 has already taken care of that problem.

We already mentioned that the two levels $\alpha = (\log\frac{1}{p})^{-1}$ and $\alpha = (\log\frac{1}{q})^{-1}$ correspond to the fill-up-level resp. to the height of the digital search tree. The precise analysis of these parameters are subtle since there is usual a $\log\log$-term involved, too. We will not discuss the details of these ranges. Technically we would have to study $F_k(s)$ for $s \to \infty$ and $s \to -\infty$.

For example, if we apply the above mentioned procedure we obtain for $\Delta_k(n)$ (in the Range 3) the asymptotic representation

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

where $G(\rho, x)$ is a periodic function and collects all contributions from the (infinitely many) saddle points.

Finally, we need to depoissonize our results. Using the depoissonzaition lemma of Jacquet-Szpankowski [3, 17] we find $\mathbf{E}X_{n,k} \sim \Delta_k(n)$, which completes the proof.

## 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 and R. Sedgewick, *Analytic Combinatorics*, Cambridge University Press, Cambridge, 2008.
[4] P. Jacquet, and W. Szpankowski, Asymptotic Behavior of the Lempel-Ziv Parsing Scheme and Digital Search Trees, *Theoretical Computer Science*, 144, 161–197, 1995.
[5] C. Knessl, and W. Szpankowski, On the Average Profile of Symmetric Digital Search Trees, preprint, 2008.
[6] D. Knuth, *The Art of Computer Programming. Sorting and Searching*, Vol. 3, Second Edition, Addison-Wesley, Reading, MA, 1998.
[7] G. Louchard, Exact and Asymptotic Distributions in Digital and Binary Search Trees, *RAIRO Theoretical Inform. Applications*, 21, 479–495, 1987.
[8] 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.
[9] 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.
[10] G. Park and W. Szpankowski, Towards a Complete Characterization of Tries, *Proc. SIAM-ACM Symposium on Discrete Algorithms* (SODA 2005), 33-42, Vancouver, 2005.
[11] G. Park Profile of Tries, Ph.D. Thesis, Purdue University, 2006.
[12] 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.
[13] B. Pittel, Asymptotic Growth of a Class of Random Trees, *Annals of Probability*, 18, 414–427, 1985.
[14] Helmut Prodinger, Digital Search Trees and Basic Hypergeometric Functions, *Bulletin of the EATCS*, 56, 1995.
[15] R. Sedgewick, *Algorithms in C: Fundamental Algorithms, Data Structures, Sorting, Searching*, Addison-Weseley, 1997.
[16] W. Szpankowski, A Characterization of Digital Search Trees From the Successful Search Viewpoint, *Theoretical Computer Science*, 85, 117–134, 1991.
[17] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, John Wiley, New York, 2001.

## Appendix A: Proof of Lemma 4.1

*Proof.* Set $\tilde{F}_k(s) = 1$ and recursively

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

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

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

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 with help of the above considerations it follows that $F_{k+1}(s) = \tilde{F}_{k+1}(s) + G(s)$, where $G(s)$ satisfies

$$G(-1) = 0 \quad \text{and} \quad G(s) - G(s-1) = 0 \quad (-k-1 < \Re(s) < 0). \tag{5.21}$$

By the above observations $G(s)$ has to be a finite linear combination of function of the form $p^{-\ell_1 s} q^{-\ell_2 s}$. However, the only periodic function of this form that meets conditions (5.21) is the zero function. Hence, $F_{k+1}(s) = \tilde{F}_{k+1}(s)$.

Now we prove (4.12). First, (4.12) is equivalent to

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

resp. to

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

We will prove this relation by induction. Of course, it is satisfied for $k = 0$. Now suppose that is holds for some $k \geq 0$. Then from (4.11) we find

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

This completes the induction proof.

Finally, since $F_k(s) = -M_k(s)/\Gamma(s)$ is analytic for $s$ with $-k-1 < \Re(s) < 0$ and $1/\Gamma(-\ell) = 0$ it also follows that $F_k(-\ell) = 0$ for $\ell = 1, 2, \ldots, k$.

## Appendix B: Proof of Lemma 4.2

*Proof.* We recall that $R_k(s) = \mathbf{A}^k[1](s)$. In particular the first few functions $R_k(s)$ are given by

$$R_0(s) = 1,$$

$$R_1(s) = \frac{p^{-s}}{1-p} + \frac{q^{-s}}{1-q},$$

$$R_2(s) = \frac{p^{-2s}}{(1-p)(1-p^2)} + \frac{p^{-s}q^{-s}}{(1-p)(1-pq)} + \frac{p^{-s}q^{-s}}{(1-q)(1-pq)} + \frac{q^{-2s}}{(1-q)(1-q^2)}.$$

With help of (4.13) we derive corresponding representations for general $k$. Recall, too, that we have assumed that $p < q$. Hence, it follows that

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

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

$$g(x,s) = \sum_{\ell\ge 0} R_\ell(s) x^\ell = \left( \sum_{\ell \ge 0} x^\ell \mathbf{A}^\ell \right) [1](s) \tag{5.22}$$

converges absolutely and represents an analytic function. We can rewrite (5.22) as

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

or as

$$(\mathbf{I} - x\mathbf{A})[g(x,\cdot)](s) = g(x,s) - x\sum_{j\ge 0} g(x,s-j)T(s-j) = 1, \tag{5.23}$$

which is the same as (4.16).

By substituting $g(x,s)$ by

$$g(s,x) = \frac{h(x,s)}{1-xT(s)}$$

in (5.23) we get a relation for $h(x,s)$ of the form

$$h(x,s) = 1 + \sum_{j\ge 1} h(x,s-j)\frac{xT(s-j)}{1-xT(s-j)}. \tag{5.24}$$

Recall that we already know that $h(x,s)$ exists for $|x| < T(\Re(s))^{-1}$. We will now use (5.24) to show that $h(x,s)$ can be analytically continued to the range $|x| < T(\Re(s)-1)^{-1}$ (and even to the range where $xT(s-m) \ne 1$) so that we also get a meromorphic continuation as proposed.

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

$$\mathbf{B}[f](s) = \sum_{j\ge 1} f(x,s-j)\frac{xT(s-j)}{1-xT(s-j)}. \tag{5.25}$$

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

$$\mathbf{B}^k[1](s) = \sum_{i_1\ge 1}\sum_{i_2\ge 1}\cdots\sum_{i_k\ge 1} U(x,s-i_1)U(x,s-i_1-i_1) \cdots U(x,s-i_1-i_2-\cdots-i_k)$$

$$= \sum_{m_k\ge k}\sum_{m_{k-1}=k-1}^{m_k-1}\sum_{m_{k-2}=k-2}^{m_{k-1}-1} \cdots \sum_{m_1=1}^{m_2-1} U(x,s-m_1)U(x,s-m_2)\cdots U(x,s-m_k).$$

Hence, we get the upper bound

$$|\mathbf{B}^k[1](s)| \le \sum_{m_k\ge k}\sum_{m_{k-1}\ge k-1}\cdots \sum_{m_1\ge 1} |U(x,s-m_1)U(x,s-m_2)\cdots U(x,s-m_k)|$$

$$= \sum_{m_1\ge 1}|U(x,s-m_1)| \cdot \sum_{m_2\ge 2}|U(x,s-m_2)| \cdots \sum_{m_k\ge k}|U(x,s-m_k)|.$$

It is clear that the series

$$S := \sum_{m\ge 1}|U(x,s-m)| = \sum_{m\ge 1}\frac{|xT(s-m)|}{|1-xT(s-m)|}$$

converges if $xT(s-m) \ne 1$ for all $m \ge 1$. Note that $T(s-m) = O(q^m)$. Thus for any choice of $x$ and $s$ there are only finitely many exceptions where $xT(s-m) = 1$. Let $k_0$ be any value with

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

Then we have for all $k \ge k_0$

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

Hence, we can set

$$h(x,s) = \sum_{k\ge 0}\mathbf{B}^k[1](s) \tag{5.26}$$

which obviously satisfies (5.24). Furthermore we have the upper bound $|h(x,s)| \le 2(2S)^{k_0}$.