External Profile of Symmetric Digital Search Trees (Extended Abstract)

Michael Drmota*
Institute for Discrete Mathematics and Geometry
Technical University of Vienna
1040 Vienna
Austria

Michael Fuchs†
Department of Applied Mathematics
National Chiao Tung University
Hsinchu 300
Taiwan

Hsien-Kuei Hwang
Institute of Statistical Science
Academia Sinica
Taipei 115
Taiwan

Ralph Neininger
Institute for Mathematics
Goethe University
60054 Frankfurt a.M.
Germany

October 11, 2016

Abstract

The external profile is among the first examined shape parameters of symmetric digital search trees in connection with the performance of unsuccessful search of a random query in the early 1970s. However, finer and important properties beyond the mean such as the variance and the limit law have remained unknown. In this extended abstract, we describe the first results for the asymptotic variance and the limit law of the external profile. In particular, the analysis of the variance turns out to be highly demanding and nontrivial, and we need diverse techniques from analytic combinatorics to unveil its asymptotic behaviors.

1 Introduction and Results

Digital trees are fundamental data structures in computer algorithms whose analysis has attracted much attention over the last four decades. One of the three main classes of digital trees is the *digital search tree* (DST for short), introduced by Coffman and Eve in 1970 [2]. While DSTs are generally less widely used as tree data structures than the other two classes of trees, which are *tries* and *Patricia tries*, their analysis is particularly related to the popular Lempel-Ziv compression scheme. Furthermore, due to the natural occurrence of differential-functional equations, their analysis is often more challenging than that of tries and PATRICIA tries.

We begin with the definition of DSTs, which are the main object of study in this paper. Similar to other digital trees, they are built from digital data consisting of records that are represented by 0-1 strings. Assume that we have n such records. Then, the digital search tree of these records is built as follows. The first record is stored in the root. All other records are distributed to the left- and the right-subtree according to whether their first bit is 0 and 1, respectively. The subtrees of the root are built according to the same rules but by using the next digit in further directing the strings to their subtrees. Note that the resulting tree is a binary tree with (internal) nodes holding the records. External nodes, which represent places where future records can be inserted, are often added to the tree (in fact, two external nodes are automatically created in the algorithmic implementation for each new internal node); see Figure 1 for an example of a DST built from 5 records (internal nodes are represented by circles and external nodes by rectangles).

For the purpose of analysis, we equip the input data with the following random model, which then yields random trees. Assume that bits in the strings are independent and identically distributed with a common Bernoulli random variable with parameter $0 < p < 1$. Throughout the paper, we fix $p = \frac{1}{2}$, namely, we consider only the symmetric case. This random model is called the *(symmetric) Bernoulli model* and the cor-

*Partially supported by the grant FWF F50-02
†Partially supported by the grants MOST-104-2923-M-009-006-MY3 and MOST-105-2115-M-009-010-MY2

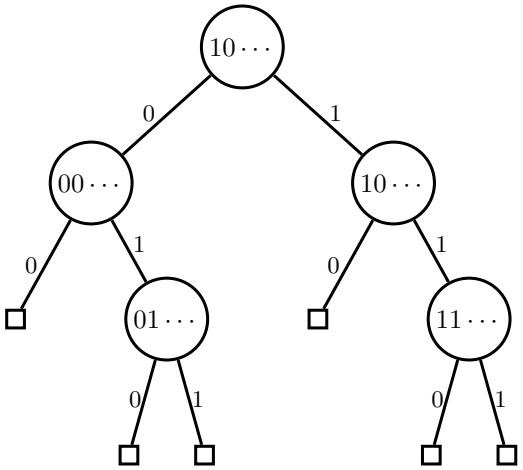

Figure 1: *A DST built from* 5 *records.*

responding random tree is called *random (symmetric) digital search tree*. It is simple yet mathematically tractable and has sufficient predictive power in general.

Under this random model, we study in this paper the external and internal profiles which are defined as follows: the external profile of a random symmetric digital search tree of size n is a double-indexed sequence of random variables $X_{n,k}$ which counts the number of external nodes at (horizontal) level k; similarly, the internal profile $I_{n,k}$ is defined (with external nodes simply replaced by internal nodes). We will only discuss in detail our results for the external profile; almost the same results hold for the internal profile (details are postponed to the journal version of this paper).

The main reason why the profiles are interesting is that they are *fine shape characteristics* encoding the complete silhouette of the tree. In particular, many other shape parameters that have been investigated since the introduction of DSTs are closely related to the profiles and can thus be analyzed in a uniform way via the profiles; these shape parameters include

- unsuccessful search U_n: the distance between a randomly chosen external node and the root; its distribution is given by the external profile divided by n; see (1.1);
- successful search or depth D_n: the distance between a randomly chosen internal node and the root; its distribution is given by the internal profile divided by n;
- (external) path length T_n: the sum of root-distances of all external nodes, or equivalently, $\sum_k kX_{n,k}$;
- height H_n: the length of the longest path from the root to an external node, or $\max\{k : X_{n,k} > 0\}$;
- fill-up level F_n: the first level from the root at which the number of internal nodes is not a power of two, or $\min\{k : X_{n,k} > 0\}$.

See for example [4, 9] and the references therein for more shape parameters in DSTs.

In fact, the external profile was one of the very first shape parameters which was analyzed for DSTs due to its close relationship with the cost of unsuccessful search; see Knuth [15] and Konheim and Newman [16]. More precisely, the distribution of unsuccessful search is obtained from the mean of the external profile as follows.

$$\mathbb{P}(U_n = k) = \frac{\mathbb{E}(X_{n,k})}{n}. \tag{1.1}$$

A similar connection between the distribution of the internal profile and the depth holds.

Despite the long history and the rich connection with other shape parameters, our understanding of the profiles of symmetric DSTs is still incomplete. Following [15, 16], Louchard [17] derived an explicit expression for the expected profile; see also [4, 5, 20, 23]. Louchard also established an asymptotic result for the mean profile in the most important range $k = \log_2 n + O(1)$ (where most nodes lie), characterizing the asymptotic distribution of unsuccessful search and the depth. These results were then extended in [4, 5, 20, 14]. We extend further the study in this paper to the variance of the profile for which a heavy analysis is carried out. We also clarify the asymptotic normality of the profiles in the range where the variance becomes unbounded.

Before stating our results, we briefly recall what is known about the profiles of asymmetric DSTs and the other two classes of digital trees. First, our results for symmetric DSTs will resemble those for symmetric tries; see Park et al. [22]. However, for symmetric tries, simple explicit expressions for the mean and the variance of the profiles are available making the resulting analysis very easy when compared to that of DST. In fact, the main contribution of [22] lies in the analysis of the profiles of asymmetric tries which turned out to be both highly non-trivial and interesting (the authors in [22] derived the mean, variance and limit laws). A similar study has been performed for the profiles of asymmetric PATRICIA tries for which the mean has been considered in [18] and more refined properties such as variance and limit laws were obtained in the very recent paper [19]. Finally, the means and variances of the profiles of asymmetric DSTs have been analyzed in [5] and in [11], respectively. Among the two

missing cases, which are symmetric PATRICIA tries and the symmetric DSTs, we address the latter in this paper.

We now state our results, focusing only on the external profile. The corresponding results for the internal profile will be given in the journal version of this paper. First, recall the following function and related sequence which have been used in most of the papers on symmetric DSTs:

$$Q(z) = \prod_{\ell\geqslant 1}\Big(1-\frac{z}{2^\ell}\Big), \quad \text{and}$$

$$Q_n = \prod_{1\leqslant \ell\leqslant n}\big(1-2^{-\ell}\big) = \frac{Q(2^{-n})}{Q(1)}.$$

Note that $\lim_{n\to\infty} Q_n$ exits and equals $Q(1) =: Q_\infty$.

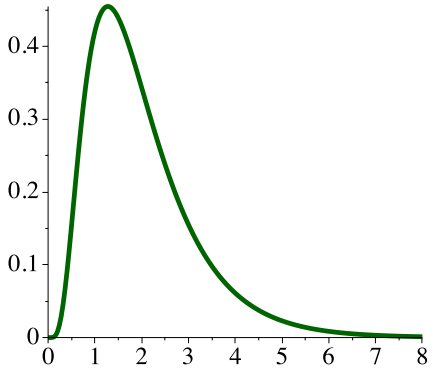

Figure 2: *The function F.*

We will first discuss the mean of the external profile for which we have the following (mostly known) result.

THEOREM 1. *The expected external profile satisfies*

$$\mathbb{E}(X_{n,k}) \begin{cases} \sim \frac{2^k}{Q_k}\big(1-2^{-k}\big)^n, & \text{if } \frac{n}{2^k}\to\infty; \\ = 2^k F\big(\frac{n}{2^k}\big) + O(1), & \text{if } \frac{n}{4^k}\to 0, \end{cases} \tag{1.2}$$

where $F(x)$ is a positive function on $[0,\infty)$ defined by

$$F(x) = \sum_{j\geqslant 0} \frac{(-1)^j 2^{-\binom{j}{2}}}{Q_j Q_\infty}\, e^{-2^j x}.$$

Note that the two ranges are overlapping; more precisely, $\frac{n}{2^k}\to\infty$ if $0\leqslant k\leqslant \log_2 n-\omega_n$ for any sequence ω_n tending to infinity with n, and $\frac{n}{4^k}\to 0$ if $k\geqslant \frac{1}{2}\log_2 n+\omega_n$.

The usefulness of the asymptotic results (1.2) depends crucially on the function $F(x)$; see Figure 2 for a plot. First, the series definition of F provides itself an asymptotic expansion for large x:

$$F(x) = \frac{e^{-x}}{Q_\infty} + O(e^{-2x}). \tag{1.3}$$

On the other hand, for small x, we have (with $\xi := \frac{1}{x\log 2}$)

$$F(x) = \frac{\xi^{\frac{1}{\log 2}}}{\sqrt{2\pi x}} \exp\left(-\frac{(\log\xi\log\xi)^2}{2\log 2} - \sum_{j\in\mathbb{Z}} c_j(\xi\log\xi)^{-\chi_j}\right) \times\left(1+O\left(\frac{(\log\log\xi)^2}{\log\xi}\right)\right), \tag{1.4}$$

where $c_0 = \frac{\log 2}{12} + \frac{\pi^2}{6\log 2}$ and $c_j = \frac{1}{2j\sinh(2j\pi/\log 2)}$ for $j\neq 0$.

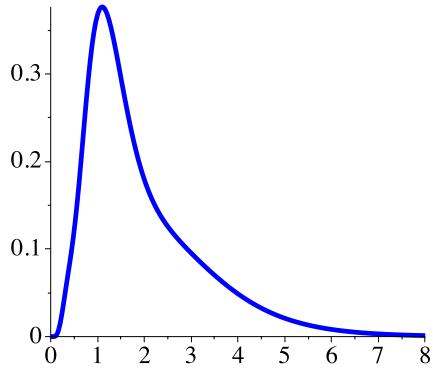

Figure 3: *The function G.*

It turns out that the variance satisfies the same types of asymptotic approximations as the mean but with the function F replaced by a much more complicated function G, which behaves similarly to F.

THEOREM 2. *The variance of the external profile satisfies*

$$\mathbb{V}(X_{n,k}) \begin{cases} \sim \frac{2^k}{Q_k}\big(1-2^{-k}\big)^n, & \text{if } \frac{n}{2^k}\to\infty; \\ = 2^k G\big(\frac{n}{2^k}\big) + O(1), & \text{if } \frac{n}{4^k}\to 0, \end{cases}$$

where $G(x)$ is a positive function on $[0,\infty)$ defined by

$$G(x) = \sum_{j,r\geqslant 0}\sum_{0\leqslant h,\ell\leqslant j} \frac{2^{-j}(-1)^{r+h+\ell} 2^{-\binom{r}{2}-\binom{h}{2}-\binom{\ell}{2}+2h+2\ell}}{Q_\infty Q_r Q_h Q_{j-h} Q_\ell Q_{j-\ell}} \times\varphi(2^{r+j}, 2^h+2^\ell; x). \tag{1.5}$$

Here

$$\varphi(u,v;x) = e^{-ux}\int_0^x te^{-(v-u)t}\,\mathrm{d}t = \begin{cases} \dfrac{e^{-ux}-((v-u)x+1)\,e^{-vx}}{(v-u)^2}, & \text{if } u\neq v; \\ \frac{1}{2}x^2e^{-ux}, & \text{if } u=v. \end{cases} \tag{1.6}$$

See Figure 3 for a plot of the function G. We can show that G satisfies the asymptotic estimates

$$G(x) \sim \begin{cases} F(x), & \text{if } x \to \infty; \\ 2F(x), & \text{if } x \to 0; \end{cases}$$

see [22] for the same type of result for symmetric tries and Devroye [3] for a general bound for the variance. A more precise approximation when $x \to \infty$ is

$$G(x) = \frac{e^{-x}}{Q_\infty} + O(xe^{-2x}),$$

where the second-order term differs from that of F; see (1.3).

The two theorems imply that the mean and the variance have very similar behaviors. In particular, they tend to infinity in the same range of k.

COROLLARY 1. *For large n, $\mathbb{E}(X_{n,k}) \to \infty$ iff $\mathbb{V}(X_{n,k}) \to \infty$.*

By Theorem 1 and the behaviors of F, we see that the range of k where the mean tends to infinity is given by

$$k_0 + \frac{\omega_n}{\log n} \leqslant k \leqslant k_1 - \frac{\omega_n}{\sqrt{\log n}}, \tag{1.7}$$

for any sequence ω_n tending to infinity with n, where

$$k_0 := \log_2 n - \log_2 \log n + \frac{\log_2 \log n}{\log n},$$

$$k_1 := \log_2 n + \sqrt{2\log_2 n} - \frac{1}{2}\log_2 \log_2 n + 1 + \frac{1}{\log 2} - \frac{3\log\log n}{4\sqrt{2(\log n)(\log 2)}}. \tag{1.8}$$

For convenience, we will refer to this range of k as the *central range*. This is also the range in which a central limit theorem holds.

THEOREM 3. *If $\mathbb{V}(X_{n,k}) \to \infty$ or k satisfies (1.7), then*

$$\frac{X_{n,k} - \mathbb{E}(X_{n,k})}{\sqrt{\mathbb{V}(X_{n,k})}} \xrightarrow{d} N(0,1).$$

Our proof relies on contraction method; see [21].

The same statement holds for the internal profile.

Due to the informativeness of the profiles, our result has many applications to other shape parameters. For brevity, we state here only a result for the height which follows from our result by the first and second moment method. This result solves a problem of Aldous and Shields [1], a heuristic solution being given previously by Knessl and Szpankowski in [12]. Recall that H_n denotes the height of random digital search tree of size n, namely, $H_n := \max\{k : X_{n,k} > 0\}$.

THEOREM 4. *Define k_H by*

$$k_H = \min\big\{k : k \geqslant \log_2 n,\ 2^k F\big(\tfrac{n}{2^k}\big) \leqslant 1\big\}.$$

Then the distribution of H_n is concentrated on k_H and $k_H - 1$:

$$\mathbb{P}(H_n = k_H \text{ or } H_n = k_H - 1) \longrightarrow 1, \qquad (n \to \infty).$$

For large n, we have $k_H = k_1 + O(1)$, where k_1 is given in (1.8).

This result is to be compared with known results for the height of tries and Patricia tries, which we summarize in Table 1 (see [6] for the height of symmetric tries, and [13] for that of Patricia tries (NB: non-rigorous proof)).

trees	$\mathbb{E}(H_n) \sim$	concentration
tries	$2\log_2 n + P(2\log_2 n)$	no
P-tries	$\log_2 n + \sqrt{2\log_2 n} + O(1)$	at 3 pts
DSTs	$\log_2 n + \sqrt{2\log_2 n}$ $-\frac{1}{2}\log_2\log_2 n + O(1)$	at 2 pts

Table 1: *A comparison of the height of random symmetric tries, Patricia tries (P-tries) and DSTs; here $P(t)$ denotes a periodic function of period* 1.

In what follows, we concentrate mainly on describing the main steps used in proving Theorem 2 for the variance, since this is the most demanding part of our analysis. For that purpose, we first show how Theorem 1 is proved, and then the proof of Theorem 2 will follow a similar pattern. The detailed proof, as well as other applications, will be contained in the journal version of this paper.

2 Asymptotics of Moments

In this section, we explain the ideas behind the proofs of Theorem 1 and Theorem 2, starting from the following distributional recurrence of the external profile

$$X_{n+1,k} \stackrel{d}{=} X_{B_n,k-1} + X^*_{n-B_n,k-1},$$

for $n \geq 0, k \geq 1$, with the initial conditions $X_{0,0} = 1, X_{0,k} = 0$ for $k \geqslant 1$, $X_{n,0} = 0$ for $n \geqslant 1$, where $B_n = \text{Binomial}\big(n, \frac{1}{2}\big)$ and $X^*_{n,k}$ is an independent copy of $X_{n,k}$.

From this recurrence, it follows that all moments satisfy the same type of recurrence of the form

$$a_{n+1,k} = 2^{1-n} \sum_{0 \leqslant j \leqslant n} \binom{n}{j} a_{j,k-1} + b_{n,k}, \tag{2.9}$$

for a given sequence $b_{n,k}$. In particular, for the expected profile $\mu_{n,k} := \mathbb{E}(X_{n,k})$, we have

$$\mu_{n+1,k} = 2^{1-n} \sum_{0\leqslant j\leqslant n} \binom{n}{j} \mu_{j,k-1}, \tag{2.10}$$

for $n \geq 0, k \geq 1$, with the initial conditions $\mu_{0,0} = 1$ and $\mu_{n,0} = \mu_{0,k} = 0$ for $n, k \geqslant 1$. Similarly, the variance $\sigma^2_{n,k} := \mathbb{V}(X_{n,k})$ satisfies (2.9) with

$$b_{n,k} = 2^{-n} \sum_{0\leqslant j\leqslant n} \binom{n}{j} \left(\mu_{n+1,k} - \mu_{j,k-1} - \mu_{n-j,k-1}\right)^2, \tag{2.11}$$

for $n \geq 0, k \geq 1$, with the initial conditions $\sigma^2_{0,0} = 1$ and $\sigma^2_{n,0} = \sigma^2_{0,k} = 0$ for $n, k \geqslant 1$. We will provide an asymptotic analysis of the solution to these recurrences.

Mean. We first use *Poissonization*, which is a standard technique in the analysis of random digital trees and operates by replacing n by a Poisson random variable with parameter z. More precisely, define the *Poisson mean* by

$$\tilde{M}_{k,1}(z) := e^{-z} \sum_{n\geqslant 0} \mu_{n,k} \frac{z^n}{n!},$$

which, by (2.10), is readily checked to satisfy the differential-functional equation

$$\tilde{M}_{k,1}(z) + \tilde{M}'_{k,1}(z) = 2\tilde{M}_{k-1,1}(z/2), \tag{2.12}$$

for $k \geq 1$, with $\tilde{M}_{0,1}(z) = e^{-z}$. This differential-functional equation can be solved by using the *Laplace transform* whose application to (2.12) yields

$$\mathscr{L}[\tilde{M}_{k,1}(z); s] = \frac{4\mathscr{L}[\tilde{M}_{k-1,1}(z); 2s]}{s+1},$$

for $k \geq 1$, with $\mathscr{L}[\tilde{M}_{0,1}(z); s] = 1/(s+1)$. Iterating k times this functional equation gives

$$\begin{aligned}\mathscr{L}[\tilde{M}_{k,1}(z); s] &= \frac{4^k}{(s+1)(2s+1)\cdots(2^k s+1)} \\ &= 2^k \sum_{0\leqslant j\leqslant k} \frac{(-1)^j 2^{-\binom{j}{2}}}{Q_j Q_{k-j}} \cdot \frac{1}{s+2^{j-k}},\end{aligned}$$

where the second equality follows from partial fraction expansion. Now, by applying the inverse Laplace transform, we obtain

$$\tilde{M}_{k,1}(z) = 2^k \sum_{0\leqslant j\leqslant k} \frac{(-1)^j 2^{-\binom{j}{2}}}{Q_j Q_{k-j}} e^{-z/2^{k-j}}.$$

A crucial observation is the following identity, which is also useful from an asymptotic point of view.

LEMMA 1. *The Poisson generating function of the expected profile satisfies*

$$\tilde{M}_{k,1}(z) = 2^k \sum_{r\geqslant 0} \frac{2^{-\binom{r+1}{2}-kr}}{Q_r} F^{(r)}\left(\frac{z}{2^k}\right),$$

where $F(z)$ is defined in Theorem 1. In particular, as $|z| \to \infty$ in the right half-plane $\Re(z) \geqslant \varepsilon$,

$$\tilde{M}_{k,1}(z) = 2^k F\left(\tfrac{z}{2^k}\right) + O(1). \tag{2.13}$$

The right-hand side is already what we anticipated for $\mu_{n,k}$ if we replace z by n (see (1.2)), and what is missing here is to justify such a replacement, or to de-Poissonize the process so as to prove the second estimate of (1.2). A general procedure to achieve this is through the use of analytic de-Poissonization techniques (essentially the saddle-point method), largely developed by Jacquet and Szpankowski [10]. In our situation, this then gives, for $\frac{n}{4^k} \to 0$,

$$\mathbb{E}(X_{n,k}) \sim \tilde{M}_{k,1}(n). \tag{2.14}$$

More precise expansions are also straightforward (referred to as the Poisson-Charlier expansion; see [9]). Note that exactly the same behavior was also established for the trie profile; see [22]. This together with (2.13) implies the claimed expansion for $\mu_{n,k}$ in Theorem 1 in the range $\frac{n}{4^k} \to 0$. The other case when $\frac{n}{2^k} \to \infty$ is much easier and can be treated by elementary means.

To complete our study of $\mu_{n,k}$ and the proof of Theorem 1, we need to clarify the asymptotic behaviors of F. The easy case is when $|z| \to \infty$ in the right half-plane $\Re(s) > 0$ in which the series definition of F is itself an asymptotic expansion. We consider the case when $|z| \to 0$ in the right half-plane using the Laplace transform:

$$\mathscr{L}[F(z); s] = \frac{1}{Q(-2s)} \qquad (\Re(s) > 1).$$

Then the asymptotic expansion (1.4) is obtained by applying the saddle-point method to the inverse Laplace integral

$$F(z) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{e^{sz}}{Q(-2s)}\,\mathrm{d}s \qquad (c > 1).$$

For this method, the asymptotic behavior of $Q(-2s)$ for large s is needed, which can be obtained by applying Mellin transform techniques because the logarithm of $Q(-2s)$ is a *harmonic sum*

$$\log Q(-2s) = \sum_{j\geqslant 0} \log\Big(1 + \frac{s}{2^j}\Big);$$

see the survey paper [7] for more similar details.

Variance. We now turn to the variance whose analysis follows the same line of arguments as for the mean but with more involved technicalities. We first introduce the Poisson generating function of the second moment:

$$\tilde{M}_{k,2}(z) := e^{-z} \sum_{n\geqslant 0} s_{n,k} \frac{z^n}{n!}.$$

Since the second moment satisfies the recurrence (2.9) with

$$b_{n,k} := 2^{1-n} \sum_{0\leqslant j\leqslant n} \binom{n}{j} \mu_{j,k-1}\mu_{n-j,k-1},$$

we then obtain the differential-functional equation

(2.15)
$$\tilde{M}_{k,2}(z) + \tilde{M}'_{k,2}(z) = 2\tilde{M}_{k-1,2}(z/2) + 2\tilde{M}_{k-1,1}(z/2)^2,$$

for $k \geq 1$, with $\tilde{M}_{0,2}(z) = e^{-z}$.

Since the binomial distribution is highly concentrated around the mean, we expect that the variance will not grow too fast when compared with the mean. This implies that there are cancellations when computing the variance from the second moment and these cancellations are often very messy to deal with. So we introduce the by now standard technique of Poissonized variance (see [8, 9]) by essentially incorporating the cancellations at the generating function level and avoiding handling the cancellations at the coefficient level. More precisely, we consider the generating function

$$\tilde{V}_k(z) := \tilde{M}_{k,2}(z) - \tilde{M}_{k,1}(z)^2 - z\tilde{M}'_{k,1}(z)^2,$$

which satisfies

$$\tilde{V}_k(z) + \tilde{V}'_k(z) = 2\tilde{V}_{k-1}(z/2) + z\tilde{M}''_{k,2}(z)^2,$$

for $k \geq 1$, with $\tilde{V}_0(z) = e^{-z} - (1+z)e^{-2z}$. By the Laplace transform and the same argument used for $\tilde{M}_{k,1}(z)$, we obtain the exact expression

(2.16)
$$\tilde{V}_k(z) = \sum_{(j,r,h,\ell)\in\mathscr{V}} \frac{2^{k-j}(-1)^{r+h+\ell} 2^{-\binom{r}{2}-\binom{h}{2}-\binom{\ell}{2}+2h+2l}}{Q_r Q_{k-j-r} Q_h Q_{j-h} Q_\ell Q_{j-\ell}} \times \varphi\Big(2^{r+j}, 2^h + 2^l; \frac{z}{2^k}\Big),$$

where

$$\mathscr{V} = \{(j,r,h,\ell) \,:\, 0 \leqslant j \leqslant k, 0 \leqslant r \leqslant k-j, 0 \leqslant h,\ell \leqslant j\}.$$

Note that $2^{r+j} = 2^h + 2^\ell$ (see (1.6)) occurs if and only if (j,r,h,ℓ) lies in the set

$$\{(j,r,h,\ell) \,:\, 1 \leqslant j \leqslant k, r = 0, h = \ell = j-1 \quad \text{or} \quad 0 \leqslant j < k, r = 1, h = \ell = j\},$$

and the corresponding terms in $\tilde{V}_k(z)$ are

$$\sum_{0\leqslant j\leqslant k} \frac{2^{k-j}2^{-2\binom{j-1}{2}+4(j-1)}}{Q_{k-j}Q_1^2Q_{j-1}^2} \cdot \frac{z^2}{2} e^{-2^j z} - \sum_{0\leqslant j<k} \frac{2^{k-j}2^{-2\binom{j}{2}+4j}}{Q_{k-j-1}Q_1Q_j^2} \cdot \frac{z^2}{2} e^{-2^{j+1}z},$$

which cancel since $Q_1 = \frac{1}{2}$. Hence, the equality part in the definition $\varphi(u,v;z)$ can be ignored. The above expression, even though much more involved than that in the Poisson mean, can still be used to derive a similar identity for $\tilde{V}_k(z)$ as in Lemma 1.

Lemma 2. *We have the identity*

$$\tilde{V}_k(z) = 2^k \sum_{m\geqslant 0} \frac{2^{-\binom{m+1}{2}-km}}{Q_m} G^{(m)}\big(\tfrac{z}{2^k}\big),$$

where G is defined in (1.5).

The next step is to clarify the asymptotic nature of this expansion.

Lemma 3. *As $|z| \to \infty$ in the right half-plane $\Re(z) \geqslant \varepsilon$, the function $\tilde{V}_k(z)$ satisfies*

$$\tilde{V}_k(z) = 2^k G\big(\tfrac{z}{2^k}\big) + O(1),$$

uniformly in z.

This provides the claimed result in Theorem 2 for the variance in the Poisson model.

Finally, from the theory developed in [9], we can replace z by n (or de-Poissonize) and get

$$\mathbb{V}(X_{n,k}) \sim \tilde{V}_k(n)$$

if $\frac{n}{4^k} \to 0$. This proves Theorem 2 for k in this range. For the other range in Theorem 2, the proof is simpler.

The final step is to characterize the asymptotics of G for large and small z. The former is not complicated and follows directly from the expression (1.5) for $G(z)$. For the latter, we again apply the Laplace transform and get

$$\mathscr{L}[G(z);s] = \sum_{j\geqslant 0} 4^{-j} \frac{\tilde{g}_j^*(2^{-j}s)}{Q(-2^{1-j}s)}$$

where

$$\tilde{g}_j^*(s) = \sum_{0\leqslant k,\ell\leqslant j} \frac{(-1)^{h+\ell} 2^{-\binom{h}{2}-\binom{\ell}{2}+2h+2\ell}}{Q_k Q_{j-k} Q_\ell Q_{j-\ell}} \times \frac{1}{(2^j s + 2^h + 2^\ell)^2}.$$

This sum is more complicated than the one (the Laplace transform of F) we encountered in the analysis of the mean, and the analysis here is expected to be more involved. Furthermore, the technique of harmonic sums we used above does not apply directly here; however, an asymptotic analysis as $|s| \to \infty$ is still possible by a more careful examination of each $g_j^*(s)$. In particular, the crucial observation is that the term $j = 2$ of the above sum is dominating.

LEMMA 4. *We have, as $|s| \to \infty$ in the right half-plane $|\Re(s)| \geqslant \varepsilon$,*

$$\frac{\tilde{g}_0^*(s)}{Q(-2s)} \sim \frac{1}{s^2 Q(-2s)}, \qquad 4^{-1}\frac{\tilde{g}_1^*(2^{-1}s)}{Q(-s)} \sim \frac{9}{sQ(-2s)}$$

and, for $j \geqslant 2$,

$$4^{-j}\frac{\tilde{g}_j^*(2^{-j}s)}{Q(-2^{1-j}s)} \sim \frac{(2j-3)!}{((j-2)!)^2}\cdot\frac{2^{\binom{j}{2}}}{s^{j-2}Q(-2s)}.$$

Thus

$$\mathscr{L}[G(z);s] \sim \frac{2}{Q(-2s)}.$$

Since the dominant term in the right-hand side equals the Laplace transform of $2F(z)$, we deduce the asymptotic estimate that $G(z) \sim 2F(z)$ when $|z| \to 0$ in the right half-plane.

References

[1] D. Aldous and P. Shields, A diffusion limit for a class of random-growing binary trees, *Probab. Theory Related Fields*, **79** (1988), 509–542.

[2] E. G. Coffman, Jr. and J. Eve, File structures using hashing functions, *Commun. ACM*, **13:7** (1970), 427–432.

[3] L. Devroye, Universal asymptotics for random tries and PATRICIA trees, *Algorithmica*, **42** (2005), 11–29.

[4] M. Drmota, *Random Trees: An Interplay between Combinatorics and Probability*, SpringerWienNewYork, Vienna, 2009.

[5] M. Drmota and W. Szpankowski, The expected profile of digital search trees, *J. Combin. Theory Ser. A*, **118** (2011), 1939–1965.

[6] P. Flajolet, On the performance evaluation of extendible hashing and trie searching, *Acta Info.* **20** (1983), 345–369.

[7] P. Flajolet, X. Gourdon, P. Dumas, Mellin transforms and asymptotics: harmonic sums, *Theoret. Comput. Sci.*, **144:1-2** (1995), 3–58.

[8] M. Fuchs, H.-K. Hwang and V. Zacharovas, An analytic approach to the asymptotic variance of trie statistics and related structures, *Theoret. Comput. Sci.* **527** (2014), 1–36.

[9] H.-K. Hwang, M. Fuchs, and V. Zacharovas, Asymptotic variance of random symmetric digital search trees, *Discrete Math. Theor. Comput. Sci. (special issue in honor of Philippe Flajolet)*, **12:2** (2010), 103–166.

[10] P. Jacquet and W. Szpankowski, Analytical de-Poissonization and its applications, *Theoret. Comput. Sci.*, **201:1-2** (1998), 1–62.

[11] R. Kazemi and M. Vahidi-Asl, The variance of the profile in digital search trees, *Discrete Math. Theor. Comput. Sci.*, **13:3** (2011), 21–38.

[12] 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. Comput.*, **30:3** (2000), 923–964.

[13] C. Knessl and W. Szpankowski, Limit laws for the height in PATRICIA tries, *J. Algorithms* **44** (2002), 63–97.

[14] C. Knessl and W. Szpankowski, On the average profile of symmetric digital search trees, *Online J. Anal. Comb.*, **4** (2009), 14 pp.

[15] D. E. Knuth, *The Art of Computer Programming, Volume 3: Sorting and Searching*, Addison-Wesley, 1973.

[16] A. G. Konheim and D. J. Newman, A note on growing binary trees, *Discrete Math.*, **4** (1973), 57-63.

[17] G. Louchard, Exact and asymptotic distributions in digital and binary search trees, *RAIRO Inform. Thér. Appl.*, **21:4** (1987), 479–495.

[18] A. Magner, C. Knessl, and W. Szpankowski, Expected external profile of PATRICIA tries, *Proceedings of the Eleventh Workshop on Analytic Algorithmics and Combinatorics (ANALCO)* (2014), 16–24.

[19] A. Magner and W. Szpankowski. Profile of PATRICIA tries, submitted.

[20] H. M. Mahmoud. Evolution of Random Search Trees, Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons Inc., New York, 1992.

[21] R. Neininger and L. Rüschendorf, A general limit theorem for recursive algorithms and combinatorial structures, *Ann. Appl. Proba.* **14** (2004), 378–418.

[22] G. Park, H.-K. Hwang, P. Nicodème and W. Szpankowski, Profile of tries, *SIAM J. Comput.*, **38:5** (2009), 1821–1880.

[23] H. Prodinger, Digital search trees and basic hypergeometric functions, *Bulletin of the EATCS*, **56** (1995), 112–115.