Discrete Mathematics and Theoretical Computer Science DMTCS vol. **12:2**, 2010, 103–166

Asymptotic variance of random symmetric digital search trees

Hsien-Kuei Hwang[1], Michael Fuchs[2] and Vytas Zacharovas[1]

[1] *Institute of Statistical Science, Academia Sinica, Taipei, 115, Taiwan*
[2] *Department of Applied Mathematics, National Chiao Tung University, Hsinchu, 300, Taiwan*

received December 30, 2009, *revised February 22, 2010*, *accepted March 1, 2010*.

Asymptotics of the variances of many cost measures in random digital search trees are often notoriously messy and involved to obtain. A new approach is proposed to facilitate such an analysis for several shape parameters on random symmetric digital search trees. Our approach starts from a more careful normalization at the level of Poisson generating functions, which then provides an asymptotically equivalent approximation to the variance in question. Several new ingredients are also introduced such as a combined use of the Laplace and Mellin transforms and a simple, mechanical technique for justifying the analytic de-Poissonization procedures involved. The methodology we develop can be easily adapted to many other problems with an underlying binomial distribution. In particular, the less expected and somewhat surprising $n(\log n)^2$-variance for certain notions of total path-length is also clarified.

Keywords: Digital search trees, Poisson generating functions, Poissonization, Laplace transform, Mellin transform, saddle-point method, Colless index, weighted path-length

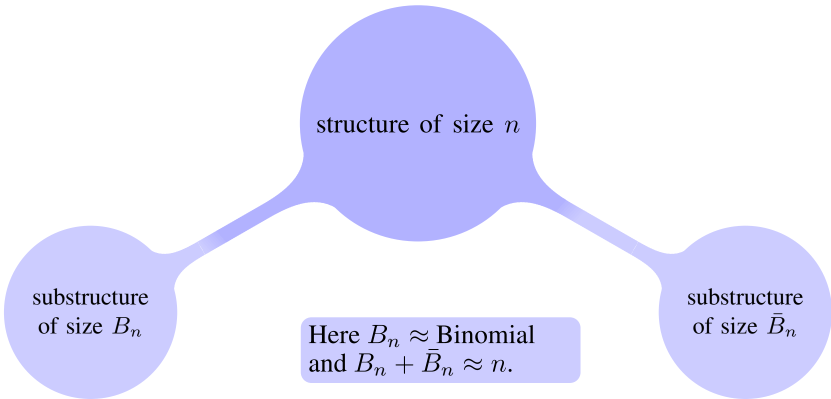
Dedicated to the 60th birthday of Philippe Flajolet

Contents

1 Introduction

The variance of a distribution provides an important measure of dispersion of the distribution and plays a crucial and, in many cases, a determinantal rôle in the limit law(i). Thus finding more effective means of computing the variance is often of considerable significance in theory and in practice. However, the calculation of the variance can be computationally or intrinsically difficult, either because of the messy procedures or cancellations involved, or because the dependence structure is too strong or simply because no simple manageable forms or reductions are available. We are concerned in this paper with random digital trees for which asymptotic approximations to the variance are often marked by heavy calculations and long, messy expressions. This paper proposes a general approach to simplify not only the analysis but also the resulting expressions, providing new insight into the methodology; furthermore, it is applicable to many other concrete situations and leads readily to discover several new results, shedding new light on the stochastic behaviors of the random splitting structures.

A binomial splitting process. The analysis of many splitting procedures in computer algorithms leads naturally to a structural decomposition (in terms of the cardinalities) of the form



where B_n is essentially a binomial distribution (up to truncation or small perturbations) and the sum of $B_n + \bar{B}_n$ is essentially n.

Concrete examples in the literature include (see the books [15, 28, 44, 50, 62] and below for more detailed references)

- tries, contention-resolution tree algorithms, initialization problem in distributed networks, and radix sort: $B_n = \text{Binomial}(n;p)$ and $\bar{B}_n = n - B_n$, namely, $\mathbb{P}(B_n = k) = \binom{n}{k}p^k q^{n-k}$ (here and throughout this paper, $q := 1-p$);
- bucket digital search trees (DSTs), directed diffusion-limited aggregation on Bethe lattice, and Eden model: $B_n = \text{Binomial}(n-b;p)$ and $\bar{B}_n = n - b - B_n$;
- Patricia tries and suffix trees: $\mathbb{P}(B_n = k) = \binom{n}{k}p^k q^{n-k}/(1-p^n-q^n)$ and $\bar{B}_n = n - B_n$.

(i) The first formal use of the term "variance" in its statistical sense is generally attributed to R. A. Fisher in his 1918 paper (see [20] or Wikipedia's webpage on variance), although its practical use in diverse scientific disciplines predated this by a few centuries (including closely-defined terms such as mean-squared errors and standard deviations).

Yet another general form arises in the analysis of multi-access broadcast channel where

$$\begin{cases} B_n = \text{Binomial}(n;p) + \text{Poisson}(\lambda), \\ \bar{B}_n = n - \text{Binomial}(n;p) + \text{Poisson}(\lambda), \end{cases}$$

see [19, 33]. For some other variants, see [2, 6, 25]. One reason of such a ubiquity of binomial distribution is simply due to the binary outcomes (either zero or one, either on or off, either positive or negative, etc.) of many practical situations, resulting in the natural adaptation of the Bernoulli distribution in the modeling.

Poisson generating function and the Poisson heuristic. A very useful, standard tool for the analysis of these binomial splitting processes is the Poisson generating function

$$\tilde{f}(z) = e^{-z} \sum_{k\ge 0} \frac{a_k}{k!} z^k,$$

where $\{a_k\}$ is a given sequence, one distinctive feature being the *Poisson heuristic*, which predicts that

$$\boxed{\textit{If } a_n \textit{ is smooth enough, then } a_n \sim \tilde{f}(n).}$$

In more precise words, if the sequence $\{a_k\}$ does not grow too fast (usually at most of polynomial growth) or does not fluctuate too violently, then a_n is well approximated by $\tilde{f}(n)$ for large n. For example, if $\tilde{f}(z) = z^m$, $m = 0, 1, \dots$, then $a_n \sim n^m$; indeed, in such a simple case, $a_n = n(n-1)\cdots(n-m+1)$.

Note that the Poisson heuristic is itself a Tauberian theorem for the Borel mean in essence; an Abelian type theorem can be found in Ramanujan's Notebooks (see [3, p. 58]).

From an elementary viewpoint, such a heuristic is based on the local limit theorem of the Poisson distribution (or essentially Stirling's formula for $n!$)

$$\frac{n^k}{k!}e^{-n} \sim \frac{e^{-x^2/2}}{\sqrt{2\pi n}}\left(1 + \frac{x^3 - 3x}{6\sqrt{n}} + \cdots\right) \qquad (k = n + x\sqrt{n}),$$

whenever $x = o(n^{1/6})$. Since a_n is smooth, we then expect that

$$\tilde{f}(n) \approx \sum_{\substack{k=n+x\sqrt{n} \\ x=O(n^{\varepsilon})}} a_k \frac{e^{-x^2/2}}{\sqrt{2\pi n}} \approx a_n \int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x = a_n.$$

On the other hand, by Cauchy's integral representation, we also have

$$\begin{aligned} a_n &= \frac{n!}{2\pi i}\oint_{|z|=n} z^{-n-1}e^z \tilde{f}(z)\,\mathrm{d}z \\ &\approx \tilde{f}(n)\frac{n!}{2\pi i}\oint_{|z|=n} z^{-n-1}e^z\,\mathrm{d}z \\ &= \tilde{f}(n), \end{aligned}$$

since the saddle-point $z = n$ of the factor $z^{-n}e^z$ is unaltered by the comparatively more smooth function $\tilde{f}(z)$.

Analytic de-Poissonization and the Poisson-Charlier expansion. The latter analytic viewpoint provides an additional advantage of obtaining an expansion by using the Taylor expansion of $\tilde{f}$ at $z = n$, yielding

$$a_n = \sum_{j\geq 0} \frac{\tilde{f}^{(j)}(n)}{j!}\tau_j(n), \tag{1}$$

where

$$\tau_j(n) := n![z^n](z-n)^j e^z = \sum_{0\leq \ell\leq j} \binom{j}{\ell}(-1)^{j-\ell}\frac{n!n^{j-\ell}}{(n-\ell)!} \qquad (j = 0, 1, \dots),$$

and $[z^n]\phi(z)$ denotes the coefficient of z^n in the Taylor expansion of $\phi(z)$. We call such an expansion *the Poisson-Charlier expansion* since the τ_j's are essentially the Charlier polynomials $C_j(\lambda, n)$ defined by

$$C_j(\lambda, n) := \lambda^{-n} n![z^n](z-1)^j e^{\lambda z},$$

so that $\tau_j(n) = n^j C_j(n, n)$. For other terms used in the literature, see [28, 29]; see also [36].

The first few terms of $\tau_j(n)$ are given as follows.

$\tau_0(n)$	$\tau_1(n)$	$\tau_2(n)$	$\tau_3(n)$	$\tau_4(n)$	$\tau_5(n)$	$\tau_6(n)$
1	0	$-n$	$2n$	$3n(n-2)$	$-4n(5n-6)$	$-5n(3n^2-26n+24)$

It is easily seen that $\tau_j(n)$ is a polynomial in n of degree $\lfloor j/2\rfloor$.

The meaning of such a Poisson-Charlier expansion becomes readily clear by the following simple but extremely useful lemma.

Lemma 1.1 *Let $\tilde{f}(z) := e^{-z}\sum_{k\geq 0} a_k z^k/k!$. If $\tilde{f}$ is an entire function, then the Poisson-Charlier expansion (1) provides an identity for a_n.*

Proof: Since $\tilde{f}$ is entire, we have

$$\sum_{n\geq 0} \frac{a_n}{n!} z^n = e^z \tilde{f}(z) = e^z \sum_{j\geq 0} \frac{\tilde{f}^{(j)}(n)}{j!}(z-n)^j,$$

and the lemma follows by absolute convergence. □

Two specific examples are worthy of mention here as they speak volume of the difference between identity and asymptotic equivalence. Take first $a_n = (-1)^n$. Then the Poisson heuristic fails since $(-1)^n \not\sim e^{-2n}$, but, by Lemma 1.1, we have the identity

$$(-1)^n = e^{-2n} \sum_{j\geq 0} \frac{(-2)^j}{j!}\,\tau_j(n).$$

See Figure 1 for a plot of the convergence of the series to $(-1)^n$.

Now if $a_n = 2^n$, then $2^n \not\sim e^n$, but we still have

$$2^n = e^n \sum_{j\geq 0} \frac{\tau_j(n)}{j!}.$$

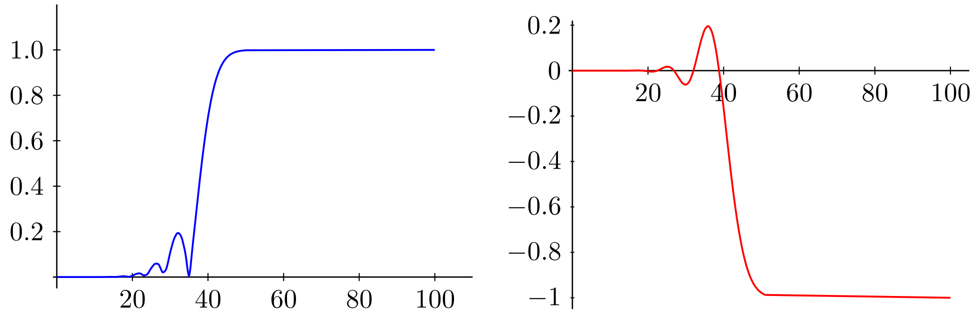


Fig. 1: *Convergence of* $e^{-2n}\sum_{j\le k}(-2)^j\tau_j(n)/j!$ *to* $(-1)^n$ *for* $n=10$ *(left) and* $n=11$ *(right) for increasing* k.

So when is the Poisson-Charlier expansion also an asymptotic expansion for a_n, in the sense that dropping all terms with $j \ge 2\ell$ introduces an error of order $\tilde{f}^{(2\ell)}n^{\ell}$ (which in typical cases is of order $\tilde{f}(n)n^{-\ell}$)? Many sufficient conditions are thoroughly discussed in Jacquet and Szpankowski's analytic de-Poissonization paper [36], although the terms in their expansions are expressed differently; see also [62].

Poissonized mean and variance. The majority of random variables analyzed in the algorithmic literature are at most of polynomial or sub-exponential (such as $e^{c(\log n)^2}$ or $e^{cn^{1/2}}$) orders, and are smooth enough. Thus the Poisson generating functions of the moments are often entire functions. The use of the Poisson-Charlier expansion is then straightforward, and in many situations it remains to justify the asymptotic nature of the expansion.

For convenience of discussion, let $\tilde{f}_m(z)$ denote the Poisson generating function of the m-th moment of the random variable in question, say X_n. Then by Lemma 1.1, we have the identity

$$\mathbb{E}(X_n) = \sum_{j\ge 0}\frac{\tilde{f}_1^{(j)}(n)}{j!}\tau_j(n),$$

and for the second moment

$$\mathbb{E}(X_n^2) = \sum_{j\ge 0}\frac{\tilde{f}_2^{(j)}(n)}{j!}\tau_j(n), \tag{2}$$

provided only that the two Poisson generating functions $\tilde{f}_1$ and $\tilde{f}_2$ are entire functions.

These identities suggest that a good approximation to the variance of X_n be given by

$$\mathbb{V}(X_n) = \mathbb{E}(X_n^2) - (\mathbb{E}(X_n))^2 \approx \tilde{f}_2(n) - \tilde{f}_1(n)^2,$$

which holds true for many cost measures, where we can indeed replace the imprecise, approximately equal symbol "≈" by the more precise, asymptotically equivalent symbol "∼". However, for a large class of problems for which the variance is essentially linear, meaning roughly that

$$\lim_{n\to\infty}\frac{\log \mathbb{V}(X_n)}{\log n} = 1, \tag{3}$$

the Poissonized variance $\tilde{f}_2(n) - \tilde{f}_1(n)^2$ is not asymptotically equivalent to the variance. This is the case for the total cost of constructing random digital search trees, for example. One technical reason is that there are additional cancellations produced by dominant terms. The next question is then: can we find a better normalized function so that the variance is asymptotically equivalent to its value at n?

Poissonized variance with correction. The crucial step of our approach that is needed when the variance is essentially linear is to consider

$$\tilde{V}(z) := \tilde{f}_2(z) - \tilde{f}_1(z)^2 - z\tilde{f}_1'(z)^2, \tag{4}$$

and it then turns out that

$$\mathbb{V}(X_n) = \tilde{V}(n) + O((\log n)^c),$$

in all cases we consider for some $c \geq 0$. The asymptotics of the variance is then reduced to that of $\tilde{V}(z)$ for large z, which satisfies, up to non-homogeneous terms, the same type of equation as $\tilde{f}_1(z)$. Thus the same tools used for analyzing the mean can be applied to $\tilde{V}(z)$.

To see how the last correction term $z\tilde{f}_1'(z)^2$ appears, we write $\tilde{D}(z) := \tilde{f}_2(z) - \tilde{f}_1(z)^2$, so that $\tilde{f}_2(z) = \tilde{D}(z) + \tilde{f}_1(z)^2$, and we obtain, by substituting this into (2),

$$\begin{aligned}
\mathbb{V}(X_n) &= \mathbb{E}(X_n^2) - (\mathbb{E}(X_n))^2 \\
&= \sum_{j\geq 0} \frac{\tilde{f}_2^{(j)}(n)}{j!}\tau_j(n) - \left(\sum_{j\geq 0} \frac{\tilde{f}_1^{(j)}(n)}{j!}\tau_j(n)\right)^2 \\
&= \tilde{D}(n) - n\tilde{f}_1'(n)^2 - \frac{n}{2}\tilde{D}''(n) + \text{smaller-order terms.}
\end{aligned}$$

Now take $\tilde{f}_1(n) \asymp n\log n$. Then the first term following $\tilde{D}(n)$ is generally not smaller than $\tilde{D}(n)$ because

$$n\tilde{f}_1'(n)^2 \asymp n(\log n)^2,$$

while $\tilde{D}(n) \asymp n(\log n)^2$, at least for the examples we discuss in this paper. Note that the variance is in such a case either of order $n\log n$ or of order n. Thus to get an asymptotically equivalent approximation to the variance, we need at least an additional correction term, which is exactly $n\tilde{f}_1'(n)^2$.

The correction term $n\tilde{f}_1'(n)^2$ already appeared in many early papers by Jacquet and Régnier (see [34]).

A viewpoint from the asymptotics of the characteristic function. Most binomial recurrences of the form

$$X_n \stackrel{d}{=} X_{B_n} + X^*_{\bar{B}_n} + T_n, \tag{5}$$

as arising from the binomial splitting processes discussed above are asymptotically normally distributed, a property partly ascribable to the highly regular behavior of the binomial distribution. Here the (X_n^*) are independent copies of the (X_n) and the random or deterministic non-homogeneous part T_n is often called the "toll-function," measuring the cost used to "conquer" the two subproblems. Such recurrences have been extensively studied in numerous papers; see [36, 52, 58, 59] and the references therein.

The correction term we introduced in (4) for Poissonized variance also appears naturally in the following heuristic, formal analysis, which can be justified when more properties are available. By definition and formal expansion

$$\begin{aligned} e^{-z}\sum_{n\geq 0}\mathbb{E}\left(e^{X_n i\theta}\right)\frac{z^n}{n!} &= \sum_{m\geq 0}\frac{\tilde{f}_m(z)}{m!}(i\theta)^m \\ &= \exp\left(\tilde{f}_1(z)i\theta - \frac{\tilde{D}(z)}{2}\theta^2 + \cdots\right), \end{aligned}$$

where $\tilde{D}(z) := \tilde{f}_2(z) - \tilde{f}_1(z)^2$, we have

$$\mathbb{E}\left(e^{(X_n - \tilde{f}_1(n))i\theta}\right) \approx \frac{n!}{2\pi i}\oint_{|z|=n} z^{-n-1}\exp\left(z + \left(\tilde{f}_1(z) - \tilde{f}_1(n)\right)i\theta - \frac{\tilde{D}(z)}{2}\theta^2 + \cdots\right)\mathrm{d}z.$$

Observe that with $z = ne^{it}$, we have the local expansion

$$ne^{it} - nit + \left(\tilde{f}_1(ne^{it}) - \tilde{f}_1(n)\right)i\theta - \frac{\tilde{D}(ne^{it})}{2}\theta^2 = n - \frac{nt^2}{2} - n\tilde{f}_1'(n)t\theta - \frac{\tilde{D}(n)}{2}\theta^2 + \cdots,$$

for small t. It follows that

$$\begin{aligned} \mathbb{E}\left(e^{(X_n - \tilde{f}_1(n))i\theta}\right) &\approx \frac{n!n^{-n}e^n}{2\pi}\exp\left(-\frac{\tilde{D}(n)}{2}\theta^2\right)\int_{-\varepsilon}^{\varepsilon}\exp\left(-\frac{nt^2}{2} - n\tilde{f}_1'(n)t\theta\right)\mathrm{d}t \\ &\sim \exp\left(-\frac{\theta^2}{2}\left(\tilde{D}(n) - n\tilde{f}_1'(n)^2\right)\right), \end{aligned}$$

by extending the integral to $\pm\infty$ and by completing the square. This again shows that $n\tilde{f}_1'(n)^2$ is the right correction term for the variance. For more precise analysis of this type, see [36].

A comparison of different approaches to the asymptotic variance. What are the advantages of the Poissonized variance with correction? In the literature, a few different approaches have been adopted for computing the asymptotics of the variance of the binomial splitting processes.

- Second moment approach: this is the most straightforward means and consists of first deriving asymptotic expansions of sufficient length for the expected value and for the second moment, then considering the difference $\mathbb{E}(X_n^2) - (\mathbb{E}(X_n))^2$, and identifying the lead terms after cancellations of dominant terms in both expansions. This approach is often computationally heavy as many terms have to be cancelled; additional complication arises from fluctuating terms, rendering the resulting expressions more messy. See below for more references.

- Poissonized variance: the asymptotics of the variance is carried out through that of $\tilde{D}(n) = \tilde{f}_2(n) - \tilde{f}_1(n)^2$. The difference between this approach and the previous one is that no asymptotics of $\tilde{f}_2(n)$ is derived or needed, and one always focuses directly on considering the equation (functional or differential) satisfied by $\tilde{D}(z)$. As we discussed above, this does not give in many cases an asymptotically equivalent estimate for the variance, because additional cancellations have to be further taken into account; see for instance [34, 35, 36].

- Characteristic function approach: similar to the formal calculations we carried out above, this approach tries to derive a more precise asymptotic approximation to the characteristic function using, say complex-analytic tools, and then to identify the right normalizing term as the variance; see the survey [36] and the papers cited there.

- Schachinger's differencing approach: a delicate, mostly elementary approach based on the recurrence satisfied by the variance was proposed in [58] (see also [59]). His approach is applicable to very general "toll-functions" T_n in (5) but at the price of less precise expressions.

The approach we use is similar to the Poissonized variance one but the difference is that the passage through $\tilde{D}(z)$ is completely avoided and we focus directly on equations satisfied by $\tilde{V}(z)$ (defined in (4)).

In contrast to Schachinger's approach, our approach, after starting from defining $\tilde{V}(z)$, is mostly analytic. It yields then more precise expansions, but more properties of T_n have to be known. The contrast here between elementary and analytic approaches is thus typical; see, for example, [7, 8]. See also Appendix for a brief sketch of the asymptotic linearity of the variance by elementary arguments.

Additional advantages that our approach offer include comparatively simpler forms for the resulting expressions, including Fourier series expansions, and general applicability (coupling with the introduction of several new techniques).

Organization of this paper. This paper is organized as follows. We start with the variance of the total path-length of random digital search trees in the next section, which was our motivating example. We then extend the consideration to bucket DSTs for which two different notions of total path-length are distinguished, which result in very different asymptotic behaviors. The application of our approach to several other shape parameters are discussed in Section 4. Table 1 summarizes the diverse behaviors exhibited by the means and the variances of the shape parameters we consider in this paper.

Shape parameters	mean	variance
Internal PL	$n \log n$	n
Key-wise PL*	$n \log n$	n
Node-wise PL*	$n \log n$	$n(\log n)^2$
Peripheral PL	n	n
$\#$(leaves)	n	n
Differential PL	n	$n \log n$
Weighted PL	$n(\log n)^{m+1}$	n

Tab. 1: *Orders of the means and the variances of all shape parameters in this paper; those marked with an * are for b-DSTs with $b \geq 2$. Here PL denotes path-length and $m \geq 0$.*

Applications of the approach we develop here to other classes of trees and structures, including tries, Patricia tries, bucket sort, contention resolution algorithms, etc., will be investigated in a future paper.

2 Digital Search Trees

We start in this section with a brief description of digital search trees (DSTs), list major shape parameters studied in the literature, and then focus on the total path-length. The approach we develop is also very useful for other linear shape measures, which is discussed in a more systematic form in the following sections.

2.1 DSTs

DSTs were first introduced by Coffman and Eve in [9] in the early 1970's under the name of sequence hash trees. They can be regarded as the bit-version of binary search trees (thus the name); see [44, p. 496 *et seq.*]. Given a sequence of binary strings, we place the first in the root node; those starting with "0" ("1") are directed to the left (right) subtree of the root, and are constructed recursively by the same procedure but with the removal of their first bits when comparisons are made. See Figure 2 for an illustration.

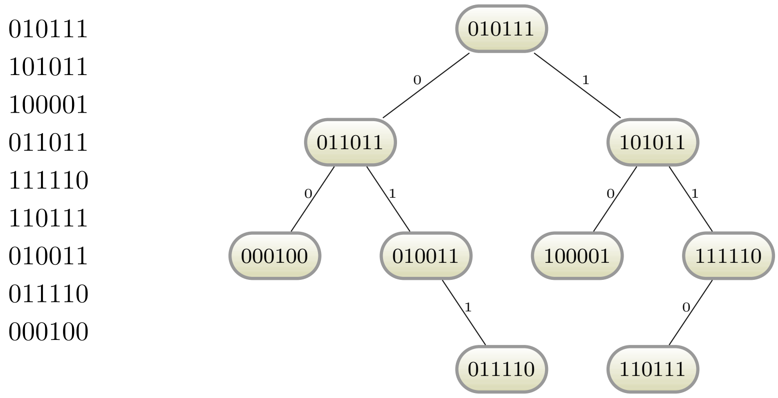


Fig. 2: *A digital search tree of nine binary strings.*

While the practical usefulness of digital search trees is limited, they represent one of the simplest, fundamental, prototype models for divide-and-conquer algorithms using coin-tossing or similar random devices. Of notable interest is its close connection to the analysis of Lempel-Ziv compression scheme that has found widespread incorporation into numerous softwares. Furthermore, the mathematical analysis is often challenging and leads to intriguing phenomena. Also the splitting mechanism of DSTs appeared naturally in a few problems in other areas; some of these are mentioned in the last section.

Random digital search trees. The simplest random model we discuss in this paper is the independent, Bernoulli model. In this model, we are given a sequence of n independent and identically distributed random variables, each comprising an infinity sequence of Bernoulli random variables with mean p, $0 < p < 1$. The DST constructed from the given random sequence of binary strings is called a *random DST*. If $p = 1/2$, the DST is said to be *symmetric*; otherwise, it is *asymmetric*. We focus on symmetric DSTs in this paper for simplicity; extension to asymmetric DSTs is possible but much harder.

Stochastic properties of many shape characteristics of random DSTs are known. Almost all of them fall into one of the two categories, according to their growth order being logarithmic or essentially linear (in the sense of (3)), which we simply refer to as "log shape measures" and "linear shape measures".

Log shape measures. The two major parameters studied in this category are *depth*, which is the distance of the root to a randomly chosen node in the tree (each with the same probability), and *height*, which counts the number of nodes from the root to one of the longest paths. Both are of logarithmic order in mean. Depth provides a good indication of the typical cost needed when inserting a new key in the tree, while height measures the worst possible cost that may be needed.

Depth was first studied in [45] in connection with the *profile*, which is the sequence of numbers, each enumerating the number of nodes with the same distance to the root. For example, the tree  has the profile $\{1, 2, 3, 2, 3\}$. For other papers on the depth of random DSTs, see [11, 12, 13, 37, 38, 39, 44, 46, 47, 50, 55, 60, 61]. The height of random DSTs is addressed in [13, 14, 43, 50, 55].

Linear shape measures. These include the total internal path-length, which sums the distance between the root and every node, and the occurrences of a given pattern (leaves or nodes satisfying certain properties); see [24, 26, 30, 31, 35, 40, 42, 44].

The profile contains generally much more information than most other shape measures, and it can to some extent be regarded as a good bridge connecting log and linear measures; see [15, 17, 45, 46] for known properties concerning expected profile of random DSTs.

Nodes of random DSTs with $p = 1/2$ are distributed in an extremely regular way, as shown in Figures 3 and 4.

2.2 Known and new results for the total internal path-length

Throughout this section, we focus on X_n, the total path length of a random digital search tree built from n binary strings. By definition and by our random assumption, X_n can be computed recursively by

$$X_{n+1} \stackrel{d}{=} X_{B_n} + X^*_{n-B_n} + n, \qquad (n \geq 0) \tag{6}$$

with the initial condition $X_0 = 0$, since removing the root results in a decrease of n for the total path length (each internal node below the root contributes 1). Here $B_n \sim \text{Binomial}(n; 1/2), X_n \stackrel{d}{=} X^*_n$, and X_n, X^*_n, B_n are independent.

Known results. It is known that (see [26, 30, 57])

$$\begin{aligned}\mathbb{E}(X_n) &= (n+1)\log_2 n + n\left(\frac{\gamma-1}{\log 2} + \frac{1}{2} - c_1 + \varpi_1(\log_2 n)\right) \\ &\quad + \frac{\gamma - 1/2}{\log 2} + \frac{5}{2} - c_1 + \varpi_2(\log_2 n) + O\left(n^{-1}\log n\right),\end{aligned} \tag{7}$$

where γ denotes Euler's constant, $c_1 := \sum_{k\geq 1}(2^k-1)^{-1}$, and $\varpi_1(t), \varpi_2(t)$ are 1-periodic functions with zero mean whose Fourier expansions are given by ($\chi_k := 2k\pi i/L$, $L := \log 2$)

$$\begin{aligned}\varpi_1(t) &= \frac{1}{L}\sum_{k\neq 0}\Gamma\left(-1-\chi_k\right)e^{2k\pi i t}, \\ \varpi_2(t) &= -\frac{1}{L}\sum_{k\neq 0}\left(1-\frac{\chi_k}{2}\right)\Gamma(-\chi_k)e^{2k\pi i t},\end{aligned} \tag{8}$$

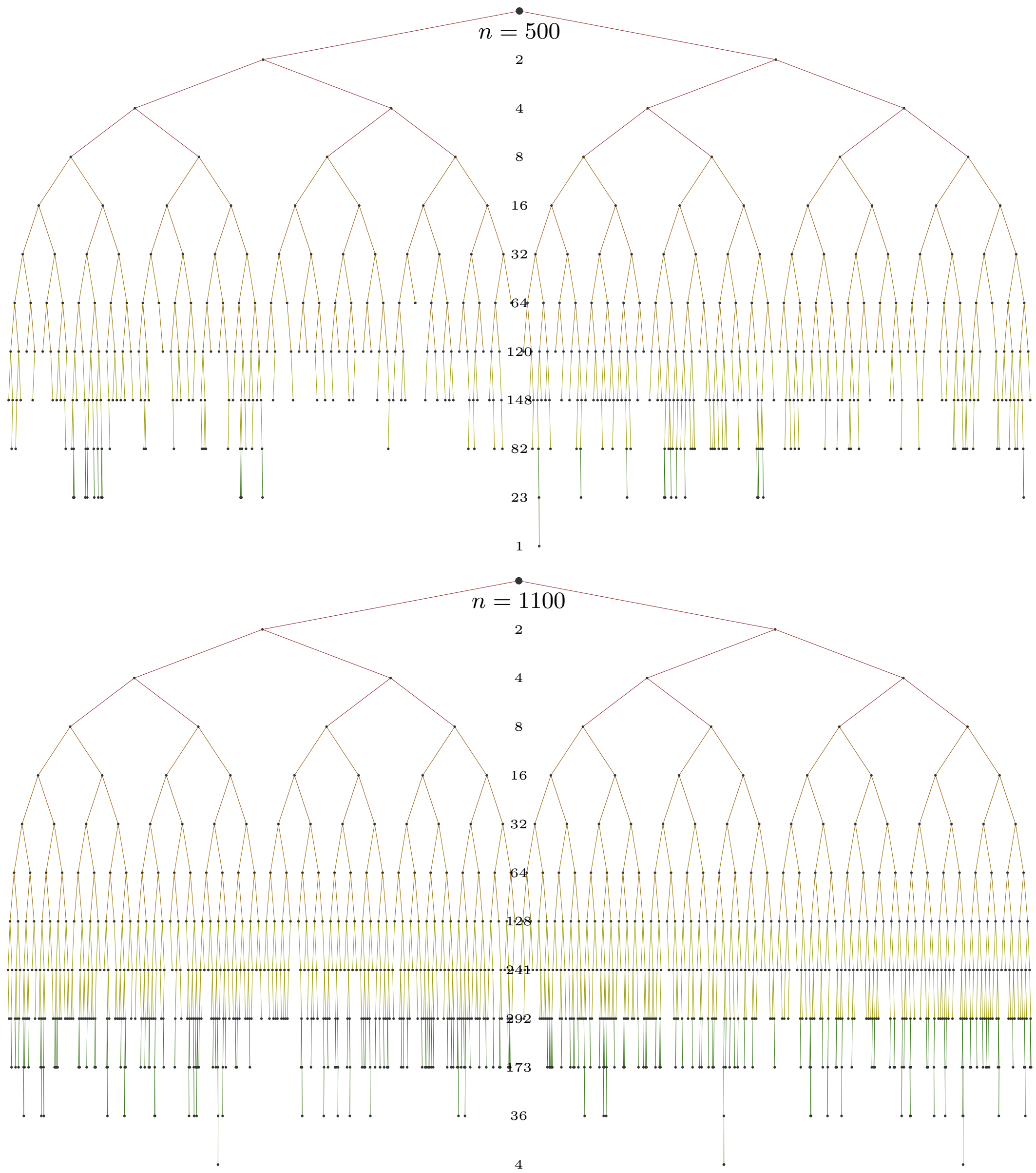


Fig. 3: *Two typical random DSTs.*

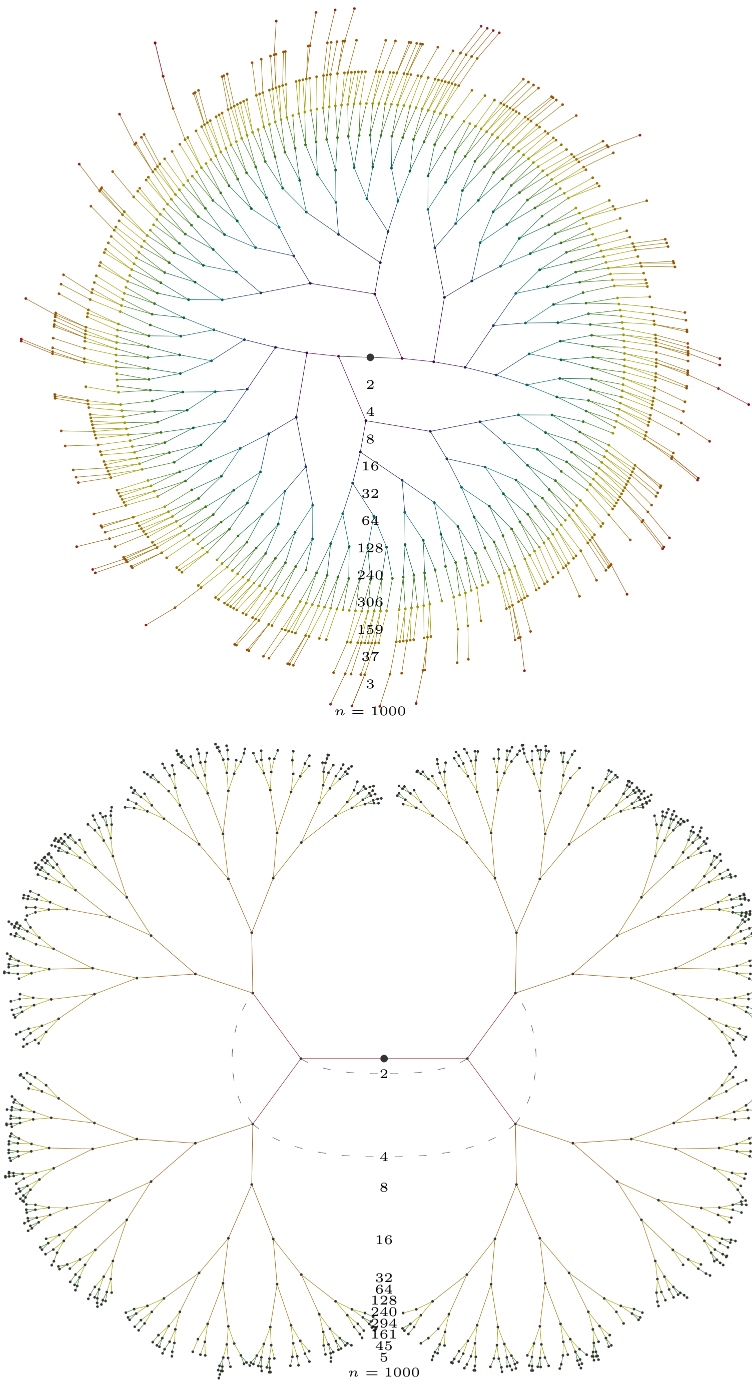


Fig. 4: *Two random DSTs of* 1000 *nodes rendered differently. For more graphical renderings of random DSTs, see the first author's webpage* `algo.stat.sinica.edu.tw`.

respectively. Here Γ denotes the Gamma function. Thus we see roughly that *random digital search trees under the unbiased Bernoulli model are highly balanced in shape*. An important feature of the periodic functions is that they are marked by very small amplitudes of fluctuation: $|\varpi_1(t)| \leq 3.4 \times 10^{-8}$ and $|\varpi_2(t)| \leq 3.4 \times 10^{-6}$. Such a quasi-flat (or smooth) behavior may in practice be very likely to lead to wrong conclusions as they are hardly visible from simulations of moderate sample sizes.

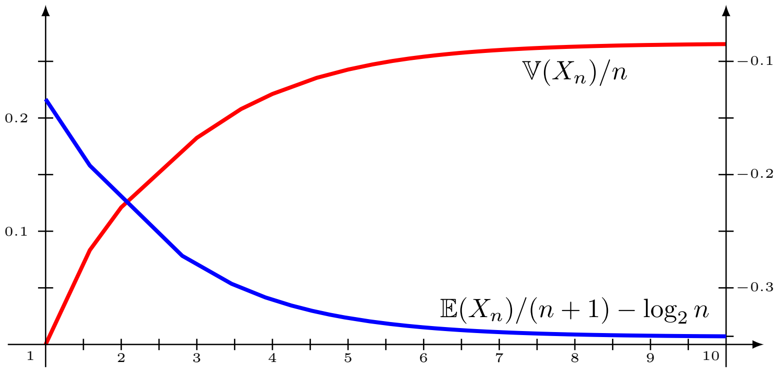


Fig. 5: *A plot of* $\mathbb{E}(X_n)/(n+1) - \log_2 n$ *in log-scale (the decreasing curve using the* y*-axis on the right-hand side), and that of* $\mathbb{V}(X_n)/n$ *in log-scale (the increasing curve using the* y*-axis on the left-hand side).*

Let

$$Q_k := \prod_{1 \leq j \leq k} \left(1 - \frac{1}{2^j}\right), \quad \text{and} \quad Q(z) := \prod_{j \geq 1} \left(1 - \frac{z}{2^j}\right). \tag{9}$$

In particular, $Q(1) = Q_\infty$. The variance was computed in [42] by a direct second-moment approach and the result is

$$\mathbb{V}(X_n) = n(C_{kps} + \varpi_{kps}(\log_2 n)) + O(\log^2 n),$$

where $\varpi_{kps}(t)$ is again a 1-periodic, zero-mean function and the mean value C_{kps} is given by ($L := \log 2$)

$$
\begin{aligned}
C_{kps} &= -\frac{28}{3L} - \frac{39}{4} + \frac{\pi^2}{2L^2} + \frac{2}{L^2} - \frac{2Q_\infty}{L} - 2\sum_{\ell\ge1}\frac{\ell 2^\ell}{(2^\ell-1)^2} + \frac{2}{L}\sum_{\ell\ge1}\frac{1}{2^\ell-1}\\
&\quad - \frac{2}{L}\sum_{\ell\ge3}\frac{(-1)^{\ell+1}(\ell-5)}{(\ell+1)\ell(\ell-1)(2^\ell-1)}\\
&\quad + \frac{2}{L}\sum_{\ell\ge1}(-1)^\ell 2^{-\binom{\ell+1}{2}}\left(\frac{L(1-2^{-\ell+1})/2-1}{1-2^{-\ell}} - \sum_{r\ge2}\frac{(-1)^{r+1}}{r(r-1)(2^{r+\ell}-1)}\right)\\
&\quad + \sum_{\ell\ge3}\sum_{2\le r<\ell}\binom{\ell+1}{r}\frac{Q_{r-2}Q_{\ell-r-1}}{2^\ell Q_\ell}\sum_{j\ge\ell+1}\frac{1}{2^j-1} - 2\left[\varpi_1^{[1]}\varpi_2^{[2]}\right]_0 - \left[(\varpi_1^{[1]})^2\right]_0\\
&\quad + 2\sum_{\ell\ge2}\frac{1}{2^\ell Q_\ell}\sum_{r\ge0}\frac{(-1)^r 2^{-\binom{r+1}{2}}}{Q_r}Q_{r+\ell-2}\times\\
&\qquad \times\Bigg\{-\sum_{j\ge1}\frac{1}{2^{j+r+\ell+2}-1}\left(2^\ell-\ell-2+\sum_{2\le i<\ell}\binom{\ell+1}{i}\frac{1}{2^{r+i-1}-1}\right)\\
&\qquad + \frac{1}{(1-2^{-\ell-r})^2} + \frac{\ell+1}{(1-2^{1-\ell-r})^2} - \frac{1}{L(1-2^{1-\ell-r})}\\
&\qquad - \sum_{2\le j\le\ell+1}\binom{\ell+1}{j}\frac{1}{2^{r+j-1}-1} + \frac{1}{L}\sum_{1\le j\le\ell+1}\binom{\ell+1}{j}\frac{1}{2^{r+j}-1}\\
&\qquad + \frac{1}{L}\sum_{0\le j\le\ell+1}\binom{\ell+1}{j}\sum_{i\ge1}\frac{(-1)^i}{(i+1)(2^{r+j+i}-1)}\Bigg\}.
\end{aligned}
$$

Here $[\varpi_1\varpi_2]_0$ denotes the mean value of the function $\varpi_1(t)\varpi_0(t)$ over the unit interval. The long expression obviously shows the complexity of the asymptotic problem.

We show that this long expression can be largely simplified. Before stating our result, we mention that the asymptotic normality of X_n (in the sense of convergence in distribution) was first proved in [35] by a complex-analytic approach; for other approaches, see [59] (martingale difference), [31] (method of moments), [52] (contraction method).

A new asymptotic approximation to $\mathbb{V}(X_n)$. Define

$$G_2(\omega) = Q_\infty \sum_{j,h,\ell\ge0}\frac{(-1)^j 2^{-\binom{j+1}{2}+j(\omega-2)}}{Q_jQ_hQ_\ell 2^{h+\ell}}\varphi(\omega; 2^{-j-h}+2^{-j-\ell}), \tag{10}$$

where for $0<\Re(\omega)<3$ and $x>0$

$$\varphi(\omega;x) := \int_0^\infty \frac{s^{\omega-1}}{(s+1)(s+x)^2}\,\mathrm{d}s,$$

which, by the relation

$$\int_0^\infty \frac{s^{\omega-1}}{s+1}\,\mathrm{d}s = \frac{\pi}{\sin(\pi\omega)} = \Gamma(\omega)\Gamma(1-\omega) \qquad (0 < \Re(\omega) < 1),$$

can be represented as

$$\varphi(\omega;x) = \begin{cases} \dfrac{\pi\left(1 + x^{\omega-2}((\omega-2)\xi + 1 - \omega\right)}{(x-1)^2\sin(\pi\omega)}, & \text{if } x \neq 1;\\ \dfrac{\pi(\omega-1)(\omega-2)}{2\sin(\pi\omega)}, & \text{if } x = 1. \end{cases}$$

The last expression provides indeed a meromorphic continuation of $\varphi(\omega;x)$ into the whole complex ω-plane whenever $x > 0$. In particular,

$$\varphi(2;x) := \begin{cases} \dfrac{x - \log x - 1}{(x-1)^2}, & \text{if } x \neq 1;\\ \dfrac{1}{2}, & \text{if } x = 1. \end{cases}$$

Theorem 2.1 *The variance of the total path-length of random DSTs of n nodes satisfies*

$$\mathbb{V}(X_n) = n(C_{kps} + \varpi_{kps}(\log_2 n)) + O(1), \tag{11}$$

where

$$C_{kps} = \frac{G_2(2)}{\log 2} = \frac{Q_\infty}{\log 2}\sum_{j,h,\ell\geq 0}\frac{(-1)^j 2^{-\binom{j+1}{2}}}{Q_jQ_hQ_\ell 2^{h+\ell}}\,\varphi(2;2^{-j-h}+2^{-j-\ell}),$$

and ϖ_{kps} has the Fourier series expansion

$$\varpi_{kps}(t) = \frac{1}{\log 2}\sum_{k\in\mathbb{Z}\setminus\{0\}}\frac{G_2(2+\chi_k)}{\Gamma(2+\chi_k)}\,e^{2k\pi it},$$

which is absolutely convergent.

One can derive more precise asymptotic expansions for $\mathbb{V}(X_n)$ by the same approach we use. We content ourselves with (11) for convenience of presentation.

Note that

$$\frac{G_2(2+\chi_k)}{\Gamma(2+\chi_k)} = \Gamma(-1-\chi_k)Q_\infty\sum_{j,h,\ell\geq 0}\frac{(-1)^j 2^{-\binom{j+1}{2}}}{Q_jQ_hQ_\ell 2^{h+\ell}}\lambda_k(2^{-j-h}+2^{-j-\ell}),$$

where

$$\lambda_k(t) := \begin{cases} \dfrac{1 - t^{\chi_k}(1+\chi_k(1-t))}{(1-t)^2}, & \text{if } t \neq 1;\\ \dfrac{\chi_k(\chi_k-1)}{2}, & \text{if } t = 1. \end{cases}$$

Thus the Fourier series is absolutely convergent by the order estimate (see [18])

$$|\Gamma(c+it)| = O\left(|t|^{c-1/2}e^{-\pi|t|/2}\right) \qquad (|t| \to \infty). \tag{12}$$

Numerically, $C_{kps} \approx 0.26600\,36454\,05936\ldots$, in accordance with that given in [42]. Also $|\varpi_{kps}(t)| \le 1.9 \times 10^{-5}$.

Sketch of our approach. Following the discussions in Introduction, we first prove that the Poisson-Charlier expansion for the mean and that for the second moment are not only identities but also asymptotic expansions. For that purpose, it proves very useful to introduce the following notion, which we term *JS-admissible functions* (following the survey paper [36] by Jacquet and Szpankowski). This is reminiscent of the classical H-admissible (due to Hayman) or HS-admissible (due to Harris and Schoenfeld) functions; see [28, §VIII.5].

Once we prove the asymptotic nature of the Poisson-Charlier expansions for the mean and the second moment, it remains, according again to the discussions in Introduction, to derive more precise asymptotics for the function $\tilde{V}$ (as defined in (4)), for which we will use first the Laplace transforms, normalize the Laplace transform properly, and then apply the Mellin transform. Such an approach will turn out to be very effective and readily applicable to more general cases such as bucket DSTs, which is discussed in details in the next section. The approach parallels closely in essence that introduced by Flajolet and Richmond in [24], which starts from the ordinary generating function, followed by an Euler transform, a proper normalization and the Mellin transform, and then conclude by singularity analysis; see also [10]. The path we take, however, offers additional operational advantages, as will be clear later. See Figure 7 for a diagrammatic illustration of the two analytic approaches.

2.3 Analytic de-Poissonization and JS-admissibility

The fundamental differential-functional equations for the analysis of random DSTs is of the form

$$\tilde{f}(z) + \tilde{f}'(z) = 2\tilde{f}(z/2) + \tilde{g}(z),$$

with suitably given initial value $f(0)$ and $\tilde{g}$. For such functions, it turns out that the asymptotic nature of the Poisson-Charlier expansions for the coefficients (or *de-Poissonization*) can be justified in a rather systematic way by the introduction of the notion of JS-admissible functions.

Here and throughout this paper, the generic symbols $\varepsilon, \varepsilon' \in (0,1)$ always represent arbitrarily small constants whose values are immaterial and may differ from one occurrence to another.

Definition 1 *An entire function $\tilde{f}$ is said to be JS-admissible, denoted by $\tilde{f} \in \mathscr{JS}$, if the following two conditions hold for $|z| \ge 1$.*

- ***(I)*** *There exist $\alpha, \beta \in \mathbb{R}$ such that uniformly for $|\arg(z)| \le \varepsilon$,*

$$\tilde{f}(z) = O\left(|z|^{\alpha}(\log_{+}|z|)^{\beta}\right),$$

where $\log_{+} x := \log(1+x)$.

- ***(O)*** *Uniformly for $\varepsilon \le |\arg(z)| \le \pi$,*

$$f(z) := e^{z}\tilde{f}(z) = O\left(e^{(1-\varepsilon')|z|}\right).$$

For convenience, we also write $\tilde{f} \in \mathscr{J\!S}_{\alpha,\beta}$ to indicate the growth order of $\tilde{f}$ inside the sector $|\arg(z)| \le \varepsilon$.

Note that if $\tilde{f}$ satisfies condition **(I)**, then, by Cauchy's integral representation for derivatives (or by Ritt's theorem; see [54, Ch. 1, § 4.3]), we have,

$$\begin{aligned}\tilde{f}^{(k)}(z) &= O\left(\oint_{|w-z|=\varepsilon|z|} \frac{|w|^{\alpha}(\log_+ |w|)^{\beta}}{|w-z|^{k+1}} |\,\mathrm{d}w|\right)\\ &= O\left(|z|^{\alpha-k}(\log_+ |z|)^{\beta}\right).\end{aligned}$$

Proposition 2.2 *Assume $\tilde{f} \in \mathscr{J\!S}_{\alpha,\beta}$. Let $f(z) := e^z\tilde{f}(z)$. Then the Poisson-Charlier expansion (1) of $f^{(n)}(0)$ is also an asymptotic expansion in the sense that*

$$\begin{aligned}a_n := f^{(n)}(0) = n![z^n]f(z) &= n![z^n]e^z\tilde{f}(z)\\ &= \sum_{0\le j<2k} \frac{\tilde{f}^{(j)}(n)}{j!}\tau_j(n) + O\left(n^{\alpha-k}(\log n)^{\beta}\right),\end{aligned}$$

for $k = 1, 2, \ldots$.

Proof: (Sketch) Starting from Cauchy's integral formula for the coefficients, the lemma follows from a standard application of the saddle-point method. Roughly, condition **(O)** guarantees that the integral over the circle with radius n and argument satisfying $\varepsilon \le |\arg(z)| \le \pi$ is negligible, while condition **(I)** implies smooth estimates for all derivatives (and thus error terms). □

The polynomial growth of condition **(I)** is sufficient for all our uses; see [36] for more general versions.

The real advantage of introducing admissibility is that it opens the possibility of developing closure properties as we now discuss.

Lemma 2.3 *Let m be a nonnegative integer and $\alpha \in (0,1)$.*

(i) $z^m, e^{-\alpha z} \in \mathscr{J\!S}$.

(ii) If $\tilde{f} \in \mathscr{J\!S}$, then $\tilde{f}(\alpha z), z^m\tilde{f} \in \mathscr{J\!S}$.

(iii) If $\tilde{f}, \tilde{g} \in \mathscr{J\!S}$, then $\tilde{f} + \tilde{g} \in \mathscr{J\!S}$.

(iv) If $\tilde{f} \in \mathscr{J\!S}$, then the product $\tilde{P}\tilde{f} \in \mathscr{J\!S}$, where $\tilde{P}$ is a polynomial of z.

(v) If $\tilde{f}, \tilde{g} \in \mathscr{J\!S}$, then $\tilde{h} \in \mathscr{J\!S}$, where $\tilde{h}(z) := \tilde{f}(\alpha z)\tilde{g}((1-\alpha)z)$.

(vi) If $\tilde{f} \in \mathscr{J\!S}$, then $\tilde{f}' \in \mathscr{J\!S}$, and thus $\tilde{f}^{(m)} \in \mathscr{J\!S}$.

Proof: Straightforward and omitted. □

Specific to our need for the analysis of DSTs is the following transfer principle.

Proposition 2.4 *Let $\tilde f(z)$ and $\tilde g(z)$ be entire functions satisfying*

$$\tilde f(z) + \tilde f'(z) = 2\tilde f(z/2) + \tilde g(z), \tag{13}$$

with $f(0) = 0$. Then

$$\tilde g \in \mathscr{J\!S} \quad \textit{if and only if} \quad \tilde f \in \mathscr{J\!S}.$$

Proof: Assume $\tilde g \in \mathscr{J\!S}$. We check first the condition **(O)** for $\tilde f$. Let $f(z) := e^z\tilde f(z)$ and $g(z) := e^z\tilde g(z)$. By (13),

$$f'(z) = 2e^{z/2}f(z/2) + g(z).$$

Consequently, since $f(0) = 0$,

$$f(z) = \int_0^z \left(2e^{t/2}f(t/2) + g(t)\right)\,\mathrm{d}t = z\int_0^1 \left(2e^{tz/2}f(tz/2) + g(tz)\right)\,\mathrm{d}t. \tag{14}$$

Now define

$$B(r) := \max_{z\in\mathcal{C}_{r,\varepsilon}} |f(z)|,$$

where

$$\mathcal{C}_{r,\varepsilon} := \{z\ :\ |z| \le r, \varepsilon \le |\arg(z)| \le \pi\}, \qquad (r \ge 0; 0 < \varepsilon < \pi/2).$$

Then, by (14), we have

$$\begin{aligned}
B(r) &\le r\int_0^1 \left(2e^{tr\cos(\varepsilon)/2}B(tr/2) + |g(tr)|\right)\,\mathrm{d}t\\
&= \int_0^r \left(2e^{t\cos(\varepsilon)/2}B(t/2) + O\left(e^{(1-\varepsilon)t}\right)\right)\,\mathrm{d}t\\
&\le Ce^{r\cos(\varepsilon)/2}B(r/2) + O\left(e^{(1-\varepsilon)r}\right),
\end{aligned}$$

where $C = 4/\cos\varepsilon > 1$. This suggests that we define a majorant function $K(r)$ of $B(r)$ by $K(r) = O(1)$ for $r \le 1$ and for $r \ge 1$

$$K(r) = Ce^{r\cos(\varepsilon)/2}K(r/2) + h(r),$$

where h is an entire function satisfying $h(r) = O(1)$ for $r \le 1$ and $h(r) = O\left(e^{(1-\varepsilon)r}\right)$ for $r \ge 1$. Let $\tilde K(r) := e^{-r\cos(\varepsilon)}K(r)$ and $\tilde h(r) := e^{-r\cos(\varepsilon)}h(r)$. Then since $\cos\varepsilon - 1 + \varepsilon > 0$ for $\varepsilon \in (0,1)$, we obtain

$$\tilde K(r) = C\tilde K(r/2) + \tilde h(r), \quad \tilde h(r) = O(1).$$

Thus if we choose $m = \lceil \log_2 r\rceil$ such that $2^m \ge r$ and iterate m times the functional equation, then we obtain the estimate

$$\begin{aligned}
\tilde K(r) &= \sum_{0\le k\le m} C^k\tilde h(r/2^k) + C^{m+1}\tilde K(r/2^{m+1})\\
&= O\left(\sum_{r/2^k>1} C^k + C^m\right)\\
&= O\left(r^{\log_2 C}\right).
\end{aligned}$$

Thus
$$B(r) = O\left(r^{\log_2 C} e^{r\cos\varepsilon}\right).$$
which establishes condition **(O)**.

Our proof for $\tilde{f}$ satisfying **(I)** proceeds in a similar manner and starts again from (14) but of the form
$$\tilde{f}(z) = z\int_0^1 e^{-(1-t)z}\left(2\tilde{f}(tz/2) + \tilde{g}(tz)\right)\,\mathrm{d}t.$$
Now, define
$$\tilde{B}(r) := \max_{z\in\mathcal{S}_{r,\varepsilon}} |\tilde{f}(z)|,$$
where
$$\mathcal{S}_{r,\varepsilon} := \{z \ : \ |z| \le r, |\arg(z)| \le \varepsilon\}, \qquad (r \ge 0; 0 < \varepsilon < \pi/2).$$
Then
$$\begin{aligned}
\tilde{B}(r) &\le r\int_0^1 e^{-(1-t)r\cos\varepsilon}\left(2\tilde{B}(tr/2) + |\tilde{g}(tr)|\right)\,\mathrm{d}t\\
&= \int_1^r \left(2e^{-(r-t)\cos\varepsilon}\tilde{B}(t/2) + O\left(e^{-(r-t)\cos\varepsilon}t^{\alpha}(\log_+ t)^{\beta}\right)\right)\,\mathrm{d}t + O(1)\\
&\le C\tilde{B}(r/2) + O\left(r^{\alpha}(\log_+ r)^{\beta} + 1\right),
\end{aligned}$$
where $C = 2/\cos\varepsilon > 2$. The same majorization argument used above for **(O)** then leads to
$$\tilde{B}(r) = \begin{cases} O(r^{\log_2 C}), & \text{if } \alpha < \log_2 C;\\ O(r^{\log_2 C}(\log_+ r)^{\beta+1}), & \text{if } \alpha = \log_2 C;\\ O\left(r^{\alpha}(\log_+ r)^{\beta}\right), & \text{if } \alpha > \log_2 C.\end{cases}$$
This proves **(I)** for $\tilde{f}$.

The necessity part follows trivially from Lemma 2.3. □

The estimates we derived of asymptotic-transfer type are indeed over-pessimistic when $1 \le \alpha \le \log_2 C$, but they are sufficient for our use. The true orders are those with $\varepsilon \to 0$, which can be proved by the Laplace-Mellin-de-Poissonization approach we use later.

Lemma 2.3 and Proposition 2.4 provide very effective tools for justifying the de-Poissonization of functions satisfying the equation (13), which is often carried out through the use of the increasing-domain argument (see [36]). The latter argument is also inductive in nature and similar to the one we are developing here, although it is less "mechanical" and less systematic.

2.4 Generating functions and integral transforms

Since our approach is purely analytic and relies heavily on generating functions, we first derive in this subsection the differential-functional equations we will be working on later. Then we apply the de-Poissonization tools we developed to the Poisson generating functions of the mean and the second moment and justify the asymptotic nature of the corresponding Poisson-Charlier expansions. Then we sketch the asymptotic tools we will follow based on the Laplace and Mellin transforms.

Generating functions. In terms of the moment generating function $M_n(y) := \mathbb{E}(e^{X_n y})$, the recurrence (6) translates into

$$M_{n+1}(y) = e^{ny}2^{-n} \sum_{0\le j\le n} \binom{n}{j} M_j(y)M_{n-j}(y), \qquad (n \ge 0), \tag{15}$$

with $M_0(y) = 1$.

Now consider the bivariate exponential generating function

$$F(z, y) := \sum_{n\ge0} \frac{M_n(y)}{n!} z^n.$$

Then by (15),

$$\frac{\partial}{\partial z} F(z, y) = F\left(\frac{e^y z}{2}, y\right)^2,$$

and the Poisson generating function $\tilde{F}(z, y) := e^{-z}F(z, y)$ satisfies the differential-functional equation

$$\tilde{F}(z, y) + \frac{\partial}{\partial z}\tilde{F}(z, y) = e^{(e^y-1)z}\tilde{F}\left(\frac{e^y z}{2}, y\right)^2, \tag{16}$$

with $\tilde{F}(0, y) = 1$. No exact solution of such a nonlinear differential equation is available; see [35] for an asymptotic approximation to $\tilde{F}$ for y near unity.

Mean and second moment. Let now

$$\tilde{F}(z, y) := \sum_{m\ge0} \frac{\tilde{f}_m(z)}{m!} y^m,$$

where $\tilde{f}_m(z)$ denotes the Poisson generating function of $\mathbb{E}(X_n^m)$. Then we deduce from (16) that

$$\tilde{f}_1(z) + \tilde{f}_1'(z) = 2\tilde{f}_1(z/2) + z, \tag{17}$$

$$\tilde{f}_2(z) + \tilde{f}_2'(z) = 2\tilde{f}_2(z/2) + 2\tilde{f}_1(z/2)^2 + 4z\tilde{f}_1(z/2) + 2z\tilde{f}_1'(z/2) + z + z^2, \tag{18}$$

with the initial conditions $\tilde{f}_1(0) = \tilde{f}_2(0) = 0$.

Proposition 2.5 *The Poisson-Charlier expansion for the mean and that for the second moment are both asymptotic expansions*

$$\mathbb{E}(X_n) = \sum_{0\le j<2k} \frac{\tilde{f}_1^{(j)}(n)}{j!}\,\tau_j(n) + O\left(n^{-k+1}\right),$$

$$\mathbb{E}(X_n^2) = \sum_{0\le j<2k} \frac{\tilde{f}_2^{(j)}(n)}{j!}\,\tau_j(n) + O\left(n^{-k+2}(\log n)^2\right),$$

for $k = 1, 2, \ldots$.

Proof: (Sketch) By Lemma 2.3 and Proposition 2.4, we see that both $\tilde{f}_1, \tilde{f}_2 \in \mathscr{J\!S}$, and thus we can apply Proposition 2.2. Indeed the proof of Proposition 2.4 provides already crude bounds for the growth order of $\tilde{f}_1, \tilde{f}_2$. The more precise estimates $\tilde{f}_1(z) \asymp |z||\log z|$ and $\tilde{f}_2(z) \asymp |z|^2|\log z|^2$ for z inside the sector $\{z : |\arg(z)| \le \varepsilon\}$ will be provided later in the next two subsections. □

An asymptotic approach based on Laplace and Mellin transforms. Once the de-Poissonization steps are justified, all that remains for the proof of Theorem 2.1 is to derive more precise asymptotic approximations to $\tilde{f}_1$ and $\tilde{V}$ (as defined in (4)). The approach we use begins with a more precise characterization of $\tilde{f}_1(z)$. Both $\tilde{f}_1$ and $\tilde{V}$ satisfy a differential-functional equation of the form

$$\tilde{f}(z) + \tilde{f}'(z) = 2\tilde{f}(z/2) + \tilde{g}(z),$$

with the initial condition $\tilde{f}(0) = 0$. To derive the asymptotics of $\tilde{f}$ for large complex z, we proceed along the following principal steps; see also [10].

***Laplace transform*:** The Laplace transform of $\tilde{f}$ satisfies

$$(s+1)\mathscr{L}[\tilde{f}; s] = 4\mathscr{L}[\tilde{f}; 2s] + \mathscr{L}[\tilde{g}; s], \tag{19}$$

which exists and defines an analytic function if $\tilde{g}$ grows at most polynomially for large $|z|$.

***Normalizing factor*:** Dividing both sides of (19) by $Q(-2s) = \prod_{j\ge0}(1 + s/2^j)$ gives a functional equation of the form

$$\bar{\mathscr{L}}[\tilde{f}; s] = 4\bar{\mathscr{L}}[\tilde{f}; 2s] + \frac{\mathscr{L}[\tilde{g}; s]}{Q(-2s)},$$

where $\bar{\mathscr{L}}[\tilde{f}; s] := \mathscr{L}[\tilde{f}; s]/Q(-s)$.

***Mellin transform*:** The Mellin transform of $\bar{\mathscr{L}}$ then satisfies

$$\mathscr{M}[\bar{\mathscr{L}}[\tilde{f}_1; s]; \omega] = \frac{1}{1 - 2^{2-\omega}}\mathscr{M}\left[\frac{\mathscr{L}[\tilde{g}; s]}{Q(-2s)}; \omega\right].$$

Inverting the process. We first derive the local behavior of $\tilde{\bar{\mathscr{L}}}[\tilde{f}; s]$ for small s by the Mellin inversion (often by calculus of residues after justification of analytic properties), and then the asymptotic behavior of $\tilde{f}(z)$ for large z is derived by the Laplace inversion, similar to singularity analysis.

2.5 Expected internal path-length of random DSTs

We consider in details in this subsection the expected value $\mu_n := \mathbb{E}(X_n)$ of the total internal path-length, paving the way for the asymptotic analysis of the variance. Starting from either the equation (17) or the recurrence

$$\mu_{n+1} = 2^{1-n} \sum_{0\le j\le n} \binom{n}{j} \mu_j + n \qquad (n \ge 0)$$

with $\mu_0 := 0$, there are several approaches to the asymptotics of μ_n. We will briefly describe the one using integral representation of finite differences (or Rice's integrals) and then present the Laplace and Mellin transforms we will use, which, as will become clear, is essentially the Flajolet-Richmond approach (see [24]).

Rice's integral representation. By (17), we have, with $\tilde{\mu}_n := n![z^n]\tilde{f}_1(z)$,

$$\tilde{\mu}_{n+1} = -\left(1 - 2^{1-n}\right)\tilde{\mu}_n \qquad (n \ge 0),$$

with $\tilde{\mu}_0 = 0$, which by iteration yields

$$\tilde{\mu}_n = (-1)^n Q_{n-2}, \quad Q_n := \prod_{1\le j\le n} \left(1 - 2^{-j}\right). \tag{20}$$

Thus by Rice's formula ([27])

$$\begin{aligned}\mu_n := \mathbb{E}(X_n) &= \sum_{2\le j\le n} \binom{n}{j} \tilde{\mu}_j \\ &= \frac{1}{2\pi i} \int_{(\frac{3}{2})} \frac{\Gamma(n+1)\Gamma(-s)}{\Gamma(n+1-s)} \cdot \frac{Q(1)}{(1-2^{1-s})Q(2^{1-s})} \,\mathrm{d}s,\end{aligned}$$

where the integration path $\int_{(c)}$ is along the vertical line with real part equal to c and Q is defined in (9). We then obtain (7) by standard arguments; see [26] or [50] for details.

This approach readily gives the approximation (7) for the mean and can be refined to obtain a full asymptotic expansion. However, its extension to the variance becomes extremely messy, as shown in [42].

Laplace transform. We first show that the asymptotics of $\tilde{f}_1(z)$ can be derived through a direct use of the Laplace and Mellin transforms, which relies on several ad hoc steps that are not easily extended. A more general procedure will be developed below.

By (17), we see that the Laplace transform of $f_1(z)$ satisfies the functional equation

$$(s+1)\mathscr{L}[\tilde{f}_1; s] = 4\mathscr{L}[\tilde{f}_1; 2s] + s^{-2}, \tag{21}$$

which exists and is analytic in $\mathbb{C} \setminus (-\infty, 0]$.

By dividing both sides by $s+1$ and by iteration, we get

$$\mathscr{L}[\tilde{f}_1; s] = \frac{1}{s^2} \sum_{j\ge 0} \frac{1}{(s+1)\cdots(2^j s+1)}. \tag{22}$$

On the other hand, from (20), we have

$$\begin{aligned}\mathscr{L}[\tilde{f}_1; s] &= \int_0^\infty e^{-sz} \sum_{n\geq 0} \frac{\tilde{\mu}_n}{n!} z^n dz \\ &= \sum_{n\geq 0} (-1)^n Q_n s^{-n-3}.\end{aligned}$$

This implies the identity

$$\sum_{n\geq 0} \frac{(-1)^n Q_n}{s^{n+1}} = \sum_{j\geq 0} \frac{1}{(s+1)\cdots(2^j s+1)}.$$

However, neither form is useful for our asymptotic purpose.

Now by partial fraction expansion, we obtain

$$\frac{1}{(s+1)\cdots(2^j s+1)} = \sum_{0\leq \ell\leq j} \frac{(-1)^{j-\ell} 2^{-\binom{j-\ell+1}{2}-\ell}}{(s+2^{-\ell}) Q_\ell Q_{j-\ell}}.$$

Thus

$$\begin{aligned}\mathscr{L}[\tilde{f}_1; s] &= \frac{1}{s^2} \sum_{j\geq 0} \sum_{0\leq \ell\leq j} \frac{(-1)^{j-\ell} 2^{-\binom{j-\ell+1}{2}-\ell}}{(s+2^{-\ell}) Q_\ell Q_{j-\ell}} \\ &= \frac{1}{s^2} \sum_{\ell\geq 0} \frac{1}{Q_\ell (2^\ell s+1)} \sum_{j\geq 0} \frac{(-1)^j 2^{-\binom{j+1}{2}}}{Q_j}.\end{aligned}$$

Note that

$$\sum_{j\geq 0} \frac{2^j s}{(s+1)\cdots(2^j s+1)} = 1.$$

By the Euler identity

$$\sum_{j\geq 0} \frac{q^{\binom{j}{2}} z^j}{(1-q)\cdots(1-q^j)} = \prod_{k\geq 0} \left(1+q^k z\right),$$

we see that

$$\sum_{j\geq 0} \frac{(-1)^j 2^{-\binom{j+1}{2}}}{Q_j} = Q(1) = Q_\infty \approx 0.28878809\ldots$$

This gives

$$\mathscr{L}[\tilde{f}_1; s] = \frac{Q_\infty}{s^2} \sum_{\ell\geq 0} \frac{1}{Q_\ell (2^\ell s+1)},$$

and then

$$\tilde{f}_1(z) = Q_\infty \sum_{\ell\geq 0} \frac{2^\ell}{Q_\ell} \left(e^{-z/2^\ell} - 1 + \frac{z}{2^\ell}\right). \tag{23}$$

Consequently,

$$\mu_n = Q_\infty \sum_{\ell \ge 0} \frac{2^\ell}{Q_\ell} \left((1-2^{-\ell})^n - 1 + 2^{-\ell} n \right).$$

Asymptotically, we have, by (23) and the identity

$$\frac{1}{Q(z)} = \prod_{j\ge 1} \frac{1}{1-z/2^j} = \sum_{\ell\ge 0} \frac{z^\ell}{Q_\ell 2^\ell} \qquad (|z|<2), \tag{24}$$

the Mellin integral representation

$$\tilde{f}_1(z) = \frac{1}{2\pi i} \int_{(-3/2)} \frac{Q(1)\Gamma(s) z^{-s}}{(1-2^{s+1})Q(2^{s+1})} \, \mathrm{d}s,$$

from which we derive the asymptotic approximation

$$\tilde{f}_1(z) = (z+1)\log_2 z + z\left(\frac{\gamma-1}{\log 2} + \frac{1}{2} - c_1 + \varpi_1(\log_2 z) \right) + O(1), \tag{25}$$

uniformly for $|z| \to \infty$ and $|\arg(z)| \le \pi/2 - \varepsilon$, where ϖ_1 is given in (8). (As usual, we use the asymptotic estimate (12) for the Gamma function.)

Laplace and Mellin transforms. We now re-do the analysis for $\tilde{f}_1(z)$ in a more general way that can be easily extended to other cases.

We again start from (21) and consider

$$\bar{\mathscr{L}}[\tilde{f}_1; s] := \frac{\mathscr{L}[\tilde{f}_1; s]}{Q(-s)},$$

where $Q(z)$ is defined in (9). Dividing both sides of (21) by $Q(-2s)$ yields

$$\bar{\mathscr{L}}[\tilde{f}_1; s] = 4\bar{\mathscr{L}}[\tilde{f}_1; 2s] + \frac{1}{Q(-2s)s^2}. \tag{26}$$

We now apply the Mellin transform. Note that we have, by the fact that $X_0 = X_1 = 0$ and the proof of Proposition 2.4,

$$\tilde{f}_1(z) = \begin{cases} O(z^2), & \text{if } z \to 0^+; \\ O(z^{1+\varepsilon}), & \text{if } z \to \infty. \end{cases}$$

Then

$$\mathscr{L}[\tilde{f}_1; s] = \begin{cases} O(s^{-2-\varepsilon}), & \text{as } s \to 0^+; \\ O(s^{-3}), & \text{as } s \to \infty. \end{cases}$$

On the other hand, by the Mellin transform,

$$\begin{aligned}\log Q(-2s) &= \sum_{j\geq 0} \log\left(1+\frac{s}{2^j}\right)\\ &= \frac{1}{2\pi i}\int_{(-\frac{1}{2})} \frac{\pi s^{-w}}{(1-2^w)w\sin \pi w}\,\mathrm{d}w\\ &= \frac{(\log s)^2}{2\log 2}+\frac{\log s}{2}+\sum_{k\in\mathbb{Z}} q_k s^{-\chi_k}+O(|s|^{-1}) \end{aligned} \tag{27}$$

uniformly for $|s|\to\infty$ and $|\arg(s)|\leq \pi-\varepsilon$, where $\chi_k := 2k\pi i/\log 2$,

$$q_0 = \frac{\log 2}{12}+\frac{\pi^2}{6\log 2}$$

and

$$q_k = \frac{1}{2k\sinh(2k\pi/\log 2)} \qquad (k\neq 0).$$

This asymptotic expansion, together with the Taylor expansion

$$Q(-2s) = 1+O(|s|), \qquad (|s|\to 0),$$

gives rise to

$$\bar{\mathscr{L}}[\tilde{f}_1;s] = \begin{cases} O(s^{-2-\varepsilon}), & \text{as } s\to 0^+;\\ O(s^{-M}), & \text{as } s\to\infty,\end{cases}$$

where $M>0$ is an arbitrary real number. Consequently, the Mellin transform of $\bar{\mathscr{L}}[\tilde{f}_1;s]$, denoted by $\mathscr{M}[\bar{\mathscr{L}};\omega]$, exists in the half-plane $\Re(\omega)\geq 2+\varepsilon$. Then by applying the Mellin transform to (26), we obtain

$$\mathscr{M}[\bar{\mathscr{L}};\omega] = \frac{G_1(\omega)}{1-2^{2-\omega}}, \qquad (\Re(\omega)>2),$$

where

$$G_1(\omega) := \int_0^\infty \frac{s^{\omega-3}}{Q(-2s)}\,\mathrm{d}s = \frac{\pi Q(2^{\omega-2})}{Q(1)\sin\pi\omega} = \frac{Q(2^{\omega-2})}{Q(1)}\,\Gamma(\omega)\Gamma(1-\omega), \tag{28}$$

for $\Re(\omega)>2$; see [24].

Inverse Mellin and inverse Laplace transforms. We can now apply successively the inverse Mellin and then Laplace transforms to derive the asymptotics of $\tilde{f}_1(z)$. Observe that $G_1(\omega)$ has a simple pole at $\omega=2$. By (28) or Proposition 5 in [22], we obtain

$$|G_1(c+it)| = O\left(e^{-(\pi-\varepsilon)|t|}\right),$$

for large $|t|$ and $c\in\mathbb{R}$. Then by the calculus of residues,

$$\bar{\mathscr{L}}[\tilde{f}_1;s] = \frac{1}{s^2}\log_2\frac{1}{s}+\frac{1}{s^2}\left(\frac{1}{2}-c_1+\frac{1}{\log 2}\sum_{k\in\mathbb{Z}\setminus\{0\}} G_1(2+\chi_k)s^{-\chi_k}\right)+O(|s|^{-1}),$$

uniformly for $|s| \to 0$ and $|\arg(s)| \le \pi - \varepsilon$. Using the expansion

$$Q(-s) = 1 + s + (|s|^2) \qquad (|s| \sim 0),$$

we see that

$$\mathscr{L}[\tilde{f}_1; s] = \frac{1+s}{s^2}\log_2\frac{1}{s} + \frac{1}{s^2}\left(\frac{1}{2} - c_1 + \frac{1}{\log 2}\sum_{k\in\mathbb{Z}\setminus\{0\}} G_1(2+\chi_k)s^{-\chi_k}\right) + O(|s|^{-1}),$$

uniformly for $|s| \to 0$ and $|\arg(s)| \le \pi - \varepsilon$.

Finally, we consider the inverse Laplace transform. The following simple result is very useful for our purposes.

Proposition 2.6 *Let $\tilde{f}(z)$ be a function whose Laplace transform exists and is analytic in $\mathbb{C} \setminus (-\infty, 0]$. Assume that*

$$\mathscr{L}[\tilde{f}; s] = \begin{cases} O\left(|s|^{-\alpha}|\log|s+1||^m\right), \\ cs^{-\omega}(-\log s)^m, \\ o(|s|^{-\alpha}|\log|s+1||^m), \end{cases} \tag{29}$$

uniformly for $|s| \to 0$ and $|\arg(s)| \le \pi - \varepsilon$, where $\alpha \in \mathbb{R}$, $\omega \in \mathbb{C}$ and $m = 0, 1, \ldots$. If $\mathscr{L}[\tilde{f}; s]$ satisfies

$$|\mathscr{L}[\tilde{f}; s]| = O\left(|s|^{-1-\varepsilon}\right), \tag{30}$$

as $|s| \to \infty$ in $|\arg(s)| \le \pi - \varepsilon$, then

$$\tilde{f}(z) = \begin{cases} O\left(|z|^{\alpha-1}(\log|z|)^m\right), \\ cz^{\omega-1}\displaystyle\sum_{0\le j\le m}\binom{m}{j}(\log z)^{m-j}\frac{\partial^j}{\partial\omega^j}\frac{1}{\Gamma(\omega)}, \\ o\left(|z|^{\alpha-1}(\log|z|)^m\right), \end{cases}$$

respectively, where the O- and o-terms hold uniformly for $|z| \to \infty$ and $|\arg(z)| \le \pi/2 - \varepsilon$.

Proof: Let $\tilde{\mathscr{L}}(s) = \mathscr{L}[\tilde{f}; s]$. Then by the inverse Laplace transform,

$$\tilde{f}(z) = \frac{1}{2\pi i}\int_{(1)} e^{zs}\tilde{\mathscr{L}}(s)\,\mathrm{d}s = \frac{1}{2\pi i}\int_{\mathcal{H}} e^{zs}\tilde{\mathscr{L}}(s)\,\mathrm{d}s,$$

where $\mathcal{H}$ is the Hankel contour consisting of the two rays $te^{\pm i\varepsilon} \pm i/|z|, -\infty < t \le 0$ and the semicircle $\exp(i\varphi)/|z|, -\pi/2 \le \varphi \le \pi/2$; see Figure 6.

Assume from now on $|z|$ is sufficiently large and lies in the sector with $|\arg(z)| \le \pi/2 - \varepsilon$. We prove only the O-case, the other two cases being similar. For simplicity, we consider only the case $m = 0$, the other cases being easily extended.

We split the above integral along $\mathcal{H}$ into two parts

$$\frac{1}{2\pi i}\int_{\mathcal{H}} e^{zs}\tilde{\mathscr{L}}(s)\,\mathrm{d}s = \frac{1}{2\pi i}\int_{\mathcal{H}_>} e^{zs}\tilde{\mathscr{L}}(s)\,\mathrm{d}s + \frac{1}{2\pi i}\int_{\mathcal{H}_\supset} e^{zs}\tilde{\mathscr{L}}(s)\,\mathrm{d}s,$$

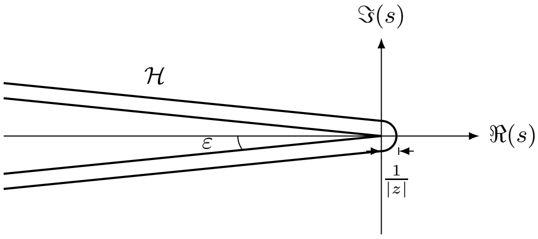


Fig. 6: *The contour $\mathcal{H}$.*

where $\mathcal{H}_>$ comprises the two rays $te^{i\varepsilon} \pm i/|z|, -\infty < t \le -T$ with $T > 1$ a fixed constant and $\mathcal{H}_\supset$ represents the remaining contour.

The integral along $\mathcal{H}_>$ is easily estimated

$$\begin{aligned}\frac{1}{2\pi i}\int_{\mathcal{H}_>} e^{zs}\tilde{\mathscr{L}}(s)\,\mathrm{d}s &= O\left(\int_{-\infty}^{-T} e^{\Re(|z|e^{i\arg(z)}(te^{i\varepsilon}+i/|z|))}|t|^{-1-\varepsilon}\,\mathrm{d}t\right)\\ &= O\left(\int_T^\infty t^{-1-\varepsilon}e^{-|z|t\cos(\arg(z)+\varepsilon)}\,\mathrm{d}t\right)\\ &= O\left(|z|^\varepsilon e^{-c|z|T}\right),\end{aligned}$$

the O-term holding uniformly for $|z| \to \infty$ provided that $|\arg(z)| + \varepsilon < \pi/2$, where $c > 0$ is a suitable constant.

For the second integral, we use (29). Then the integral along the semicircle is bounded as follows.

$$\frac{1}{2\pi|z|}\int_{-\pi/2}^{\pi/2} e^{ze^{i\theta}/|z|+i\theta}\tilde{\mathscr{L}}(e^{i\theta}/|z|)\,\mathrm{d}\theta = O\left(|z|^{\alpha-1}\right),$$

uniformly for $|z| \to \infty$. For the remaining part $t \pm i/|z|, -T < t \le 0$, we have

$$\begin{aligned}\frac{1}{2\pi i}\int_{-T}^{0} e^{z(t\pm i/|z|)}\tilde{\mathscr{L}}(t\pm i/|z|)\,\mathrm{d}t &= O\left(|z|^\alpha\int_{-T}^0 \frac{e^{c|z|t}}{(|z|^2t^2+1)^{\alpha/2}}\,\mathrm{d}t\right)\\ &= O\left(|z|^{\alpha-1}\int_0^\infty \frac{e^{-cu}}{(u^2+1)^{\alpha/2}}\,\mathrm{d}u\right)\\ &= O\left(|z|^{\alpha-1}\right),\end{aligned}$$

uniformly for $|z| \to \infty$, where $c > 0$ is a suitable constant. This completes the proof. □

Note that the inverse Laplace transform of $s^{-2}\log(1/s)$ is $z\log z - (1-\gamma)z$. This, together with a combined use of Proposition 2.6, leads to (25).

The justification of the estimate (30) is easily performed by using the relation (31) below.

The Flajolet-Richmond approach [24]. Instead of the Poisson generating function, this approach starts from the ordinary generating function $A(z) := \sum_n \mu_n z^n$.

- Then the Euler transform(ii)

  $$\hat{A}(s) := \frac{1}{s+1} A\left(\frac{1}{s+1}\right)$$

  satisfies

  $$(s+1)\hat{A}(s) = 4\hat{A}(2s) + s^{-2},$$

  identical to (21).

- The normalized function $\bar{A}(s) := \hat{A}(s)/Q(-s)$ satisfies

  $$\bar{A}(s) = 4\bar{A}(2s) + \frac{1}{s^2 Q(-2s)},$$

  again identical to (26).

- The Mellin transform of $\bar{A}$ satisfies ($\Re(\omega) > 2$)

  $$\mathscr{M}[\bar{A}; \omega] = \frac{G_1(\omega)}{1 - 2^{2-\omega}},$$

  where $G_1(\omega)$ is as defined in (28).

Then invert the process by considering first the Mellin inversion, deriving asymptotics of

$$\bar{A}(s) = \frac{1}{2\pi i} \int_{(5/2)} s^{-\omega} \frac{G_1(\omega)}{1 - 2^{2-\omega}} \,\mathrm{d}\omega,$$

as $s \to 0$ in $\mathbb{C}$. Then deduce asymptotics of

$$A(z) = \frac{1}{z} \hat{A}\left(\frac{1}{z} - 1\right),$$

as $z \to 1$. Finally, apply singularity analysis (see [23]) to conclude the asymptotics of μ_n.

The crucial reason why the two approaches are identical at certain steps is that the Laplace transform of a Poisson generating function is essentially equal to the Euler transform of an ordinary generating function; or formally,

$$\begin{aligned} \int_0^\infty e^{-sz} \sum_{n\geq 0} \frac{a_n}{n!} z^n \,\mathrm{d}z &= \sum_{n\geq 0} a_n (s+1)^{-n-1} \\ &= \frac{1}{s+1} A\left(\frac{1}{s+1}\right). \end{aligned} \tag{31}$$

(ii) For a better comparison with the approach we use, our $\hat{A}$ differs from the usual Euler transform by a factor of s.

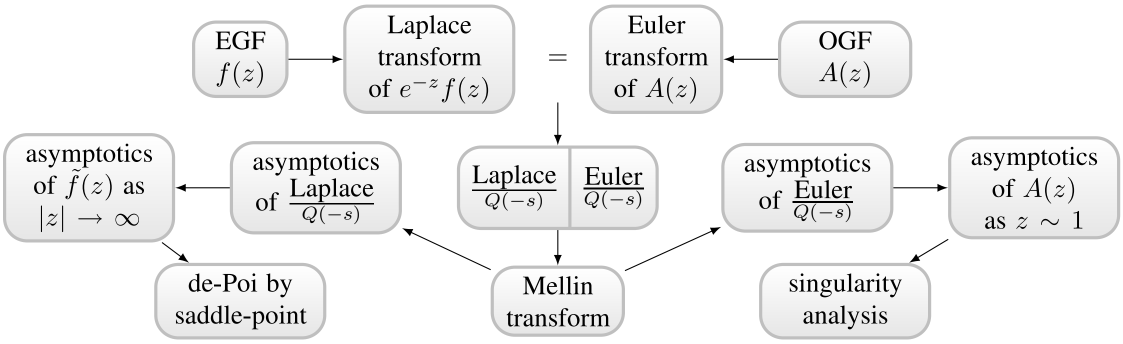


Fig. 7: *A diagrammatic comparison of the major steps used in the Laplace-Mellin (left-half) approach and the Flajolet-Richmond (right-half) approach. Here EGF denotes "exponential generating function", OGF stands for "ordinary generating function" and de-Poi is the abbreviation for de-Poissonization.*

Thus the simple result in Proposition 2.6 closely parallels that in singularity analysis. While identical at certain steps, the two approaches diverge in their final treatment of the coefficients, and the distinction here is typically that between the saddle-point method and the singularity analysis, a situation reminiscent of the use before and after Lagrange's inversion formula; see for instance [28].

The relation (31) implies that the order estimate (30) for the Laplace transform at infinity can be easily justified for all the generating functions we consider in this paper since $A(0) = 0$, implying that $A(z) = O(|z|)$ as $|z| \to 0$.

This comparison also suggests the possibility of developing de-Poissonization tools by singularity analysis, which will be investigated in details elsewhere.

2.6 Variance of the internal path-length

In this section, we apply the Laplace-Mellin-de-Poissonization approach to the Poissonized variance with correction

$$\tilde V(z) := \tilde f_2(z) - \tilde f_1(z)^2 - z\tilde f_1'(z)^2,$$

aiming at proving Theorem 2.1. The starting point of focusing on $\tilde V$ instead of on $\tilde f_2$ removes all heavy cancellations involved when dealing with the variance, a key step differing from all previous approaches.

Laplace and Mellin transform. The following lemma will be useful.

Lemma 2.7 *If*

$$\begin{cases} \tilde f_1(z) + \tilde f_1'(z) = 2\tilde f_1(z/2) + \tilde h_1(z), \\ \tilde f_2(z) + \tilde f_2'(z) = 2\tilde f_2(z/2) + \tilde h_2(z), \end{cases}$$

where all functions are entire with $\tilde f_1(0) = \tilde f_2(0) = 0$, then the function $\tilde V(z) := \tilde f_2(z) - \tilde f_1(z)^2 - z\tilde f_1'(z)^2$ satisfies

$$\tilde V(z) + \tilde V'(z) = 2\tilde V(z/2) + \tilde g(z),$$

with $\tilde V(0) = 0$, where

$$\tilde g(z) = z\tilde f_1''(z)^2 + \tilde h_2(z) - \tilde h_1(z)^2 - z\tilde h_1'(z)^2 - 4\tilde h_1(z)\tilde f_1(z/2) - 2z\tilde h_1'(z)\tilde f_1'(z/2) - 2\tilde f_1(z/2)^2.$$

Proof: Straightforward and omitted. □

By using the differential-functional equations (17) and (18) for $\tilde{f}_1(z)$ and $\tilde{f}_2(z)$, we see, by Lemma 2.7, that

$$\tilde{V}(z) + \tilde{V}'(z) = 2\tilde{V}(z/2) + z\tilde{f}_1''(z)^2, \tag{32}$$

with $\tilde{V}(0) = 0$.

Before applying the integral transforms, we need rough estimates of $\tilde{V}(z)$ near $z = 0$ and $z = \infty$. We have

$$\tilde{V}(z) = \begin{cases} O\left(z^2\right), & \text{as } z \to 0^+; \\ O(z^{1+\varepsilon}), & \text{as } z \to \infty. \end{cases} \tag{33}$$

These estimates follow from

$$z\tilde{f}_1''(z)^2 = \begin{cases} O(|z|), & \text{as } |z| \to 0; \\ O(|z|^{-1}), & \text{as } |z| \to \infty, \end{cases} \tag{34}$$

which in turn result from $X_0 = X_1 = 0$ and (25) (by the proof of condition **(I)** of Proposition 2.4). Indeed, the proof there shows that the same bounds hold uniformly for $z \in \mathbb{C}$ with $|\arg(z)| \le \pi/2 - \varepsilon$.

We now apply the Laplace transform to both sides of (32). First, observe that the Laplace transform of $\tilde{V}(z)$ exists and is analytic in $\mathbb{C} \setminus (-\infty, 0]$. Then, by (32),

$$(s+1)\mathscr{L}[\tilde{V}; s] = 4\mathscr{L}[\tilde{V}; 2s] + \tilde{g}^\star(s),$$

where $\tilde{g}^\star(s) := \mathscr{L}[z\tilde{f}_1''; s]$. Next the normalized Laplace transform

$$\bar{\mathscr{L}}[\tilde{V}; s] := \frac{\mathscr{L}[\tilde{V}; s]}{Q(-s)}$$

satisfies

$$\bar{\mathscr{L}}[\tilde{V}; s] = 4\bar{\mathscr{L}}[\tilde{V}; 2s] + \frac{\tilde{g}^\star(s)}{Q(-2s)}.$$

By (33), we obtain

$$\mathscr{L}[\tilde{V}; s] = \begin{cases} O(s^{-2-\varepsilon}), & \text{as } s \to 0^+; \\ O(s^{-3}), & \text{as } s \to \infty. \end{cases}$$

From this and the asymptotic expansion (27) of $Q(-2s)$, it follows that the Mellin transform of $\bar{\mathscr{L}}[\tilde{V}; s]$ exists in the half-plane $\Re(\omega) \ge 2 + \varepsilon$. Consequently,

$$\mathscr{M}[\bar{\mathscr{L}}[\tilde{V}; s]; \omega] = \frac{G_2(\omega)}{1 - 2^{2-\omega}}, \qquad (\Re(\omega) > 2),$$

where

$$G_2(\omega) := \mathscr{M}\left[\frac{\tilde{g}^\star(s)}{Q(-2s)}; \omega\right] = \int_0^\infty \frac{s^{\omega-1}}{Q(-2s)} \int_0^\infty e^{-zs} z\tilde{f}_1''(z)^2 \,\mathrm{d}z \,\mathrm{d}s. \tag{35}$$

By (23), we have

$$z\tilde{f}_1''(z)^2 = Q_\infty^2 \sum_{h,\ell\geq 0} \frac{1}{Q_h Q_\ell 2^{h+\ell}} z e^{-z/2^h - z/2^\ell}.$$

Substituting this and the partial fraction expansion

$$\frac{1}{Q(-2s)} = \frac{1}{Q_\infty} \sum_{j\geq 0} \frac{(-1)^j 2^{-\binom{j}{2}}}{Q_j(s+2^{-j})},$$

into (35), we obtain (10).

Inverse Mellin and inverse Laplace transforms. For the Mellin inversion, we need more precise analytic properties of $G_2(\omega)$. By (34), we deduce that the Laplace transform $\tilde{g}^\star(s)$ of $z\tilde{f}_1''(z)^2$ satisfies

$$\tilde{g}^\star(s) = \begin{cases} O(|\log s|), & \text{as } |s| \to 0; \\ O(|s|^{-2}), & \text{as } |s| \to \infty \end{cases}$$

uniformly in the cone $|\arg(s)| \leq \pi - \varepsilon$. Thus, by the asymptotic expansion (27) for $Q(-2s)$ and Proposition 5 in [22], we have

$$|G_2(c+it)| = O\left(e^{-(\pi-\varepsilon)|t|}\right),$$

for large $|t|$ and $c > 0$. Also the Mellin transform G_2 of $\tilde{g}^\star(s)/Q(-2s)$ exists in the half-plane $\Re(\omega) > 0$. Consequently, by standard calculus of residues,

$$\mathscr{L}[\tilde{V}; s] = \frac{1}{\log 2} \sum_{k\in\mathbb{Z}} G_2(2+\chi_k) s^{-2-\chi_k} + O(|s|^{-\varepsilon}),$$

uniformly for $|s| \to 0$ and $|\arg(s)| \leq \pi - \varepsilon$. This in turn yields the following expansion for $\mathscr{L}[\tilde{V}; s]$

$$\mathscr{L}[\tilde{V}; s] = \frac{1}{\log 2} \sum_{k\in\mathbb{Z}} G_2(2+\chi_k) s^{-2-\chi_k} + \frac{1}{\log 2} \sum_{k\in\mathbb{Z}} G_2(2+\chi_k) s^{-1-\chi_k} + O(|s|^{-\varepsilon}),$$

again uniformly for $|s| \to 0$ and $|\arg(s)| \leq \pi - \varepsilon$.

Finally, standard Laplace inversion gives

$$\tilde{V}(z) = \frac{z}{\log 2} \sum_{k\in\mathbb{Z}} \frac{G_2(2+\chi_k)}{\Gamma(2+\chi_k)} z^{\chi_k} + \frac{1}{\log 2} \sum_{k\in\mathbb{Z}} \frac{G_2(1+\chi_k)}{\Gamma(1+\chi_k)} z^{\chi_k} + O(|z|^{\varepsilon-1}), \tag{36}$$

uniformly for $|z| \to \infty$ and $|\arg(z)| \leq \pi/2 - \varepsilon$.

Since $\tilde{f}_2(z) = \tilde{V}(z) + \tilde{f}_1(z)^2 + z\tilde{f}_1'(z)^2$, we see from (36) and (25) that

$$\tilde{f}_2(z) \asymp \tilde{f}_1(z)^2 \asymp |z|^2 \log^2 |z| \qquad (|\arg(z)| \leq \pi/2 - \varepsilon).$$

This proves Proposition 2.5 and Theorem 2.1 by straightforward expansion. More refined calculations give

$$\mathbb{V}(X_n) = \tilde{V}(n) - \frac{n}{2}\tilde{V}''(n) - \frac{n^2}{2}\tilde{f}_1''(n)^2 + O(n^{-1}),$$

the two terms following $\tilde{V}(n)$ being both $O(1)$ and periodic in nature. It is possible to further extend the same idea and derive a full asymptotic expansion, which has also its identity nature; details will be presented in a future paper.

3 Bucket Digital Search Trees

In this section, we extend the same approach to bucket digital search trees (b-DSTs) in which each node can hold up to b keys. The construction rule is the same as DSTs, except that keys keep staying in a node as long as its capacity remains less than b; see Figure 8 for a simple example with $b = 2$. DSTs correspond to $b = 1$.

Note that when $b \geq 2$ we can distinguish two different types of total path-length: the total path-length of all keys (summing the distance between each key to the root over all keys), which will be referred to as the *total key-wise path-length* (KPL) and the total path-length of all nodes (summing the distance between each node to the root over all nodes, regardless of the number of keys in each node), referred to as the *total node-wise path-length* (NPL). When $b = 1$ the two total path-lengths coincide. For simplicity, we will use KPL and NPL, dropping the collective adjective "total". While the expected values of both TPLs are of order $n \log n$ under the same independent Bernoulli model, their variances surprisingly turn out to exhibit very different behavior; see Table 1.

010111
101011
100001
011011
111110
110111
010011
011110
000100

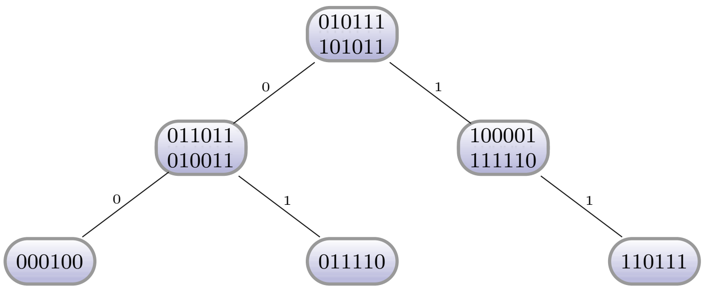


Fig. 8: *A 2-DST with nine keys. The total key-wise path-length is equal to* $4 \times 1 + 3 \times 2 = 10$ *and the total node-wise path-length equals* $2 \times 1 + 3 \times 2 = 8$.

3.1 Key-wise path-length (KPL)

We assume the same independent Bernoulli model for the input strings. Let X_n denote the KPL in a random b-DST built from n random stings. Then by definition and the independence model assumption

$$X_{n+b} \stackrel{d}{=} X_{B_n} + X^*_{n-B_n} + n, \qquad (n \geq 0) \tag{37}$$

with the initial conditions $X_0 = \cdots = X_{b-1} = 0$. Here $B_n \sim \text{Binomial}(n, 1/2)$, $X_n \stackrel{d}{=} X^*_n$, and X_n, X^*_n, B_n are independent.

Known and new results. Hubalek [30] showed, by the Flajolet-Richmond approach, that the mean satisfies

$$\mathbb{E}(X_n) = (n+b)\log_2 n + n\left(c_2 + \varpi_3(\log_2 n)\right) + c_3 + \varpi_4(\log_2 n) + O\left(n^{-1}\log n\right),$$

where c_2, c_3 are effectively computable constants and ϖ_3 and ϖ_4 are very smooth periodic functions. He also proved that the variance is asymptotically linear

$$\mathbb{V}(X_n) = n\,(C_h + \varpi_h(\log_2 n)) + O((\log n)^2),$$

where C_h is expressed in terms of a very long, involved expression and ϖ_h is a periodic function.

We improve this estimate by deriving a much simpler expression for the periodic function, including its average value C_h. To state our result, we define the following functions. Let

$$\begin{aligned}\tilde{g}(z) := {} & \left(\sum_{0\le j\le b}\binom{b}{j}\tilde{f}_1^{(j)}(z)\right)^2 + z\left(\sum_{0\le j\le b}\binom{b}{j}\tilde{f}_1^{(j+1)}(z)\right)^2 \\ & - \sum_{0\le j\le b}\binom{b}{j}\left(\left(\tilde{f}_1^2(z)\right)^{(j)} + \left(z\tilde{f}_1'(z)^2\right)^{(j)}\right).\end{aligned} \tag{38}$$

It is easily seen that $\tilde{g}(z)$ is of the form

$$\tilde{g}(z) = \sum_{2\le i_1,i_2\le b}\tilde{g}_{i_1,i_2}\tilde{f}_1^{(i_1)}(z)\tilde{f}_1^{(i_2)}(z) + z\sum_{2\le i_1,i_2\le b+1}\tilde{g}'_{i_1,i_2}\tilde{f}_1^{(i_1)}(z)\tilde{f}_1^{(i_2)}(z), \tag{39}$$

where $\tilde{g}_{i_1,i_2}, \tilde{g}'_{i_1,i_2} \ge 0$ are given explicitly by

$$\begin{aligned}\tilde{g}_{i_1,i_2} &= \binom{b}{i_1}\binom{b}{i_2} - \binom{b}{i_1}\binom{b-i_1}{i_2} - (b-i_1+1)\binom{b}{i_1-1}\binom{b-i_1}{i_2-1},\\ \tilde{g}'_{i_1,i_2} &= \binom{b}{i_1-1}\binom{b}{i_2-1} - \binom{b}{i_1-1}\binom{b-i_1+1}{i_2-1},\end{aligned}$$

both coefficients being symmetric in i_1 and i_2. Define

$$G_2(\omega) = \int_0^\infty \frac{s^{\omega-1}}{Q(-2s)^b}\int_0^\infty e^{-zs}\tilde{g}(z)\,\mathrm{d}z\,\mathrm{d}s,$$

which is well-defined for $\Re(\omega) > 0$, as we will see later.

Theorem 3.1 *The variance of the total key-wise path-length of random b-DSTs of n strings satisfies*

$$\mathbb{V}(X_n) = n\,(C_h + \varpi_h(\log_2 n)) + O(1), \tag{40}$$

where

$$C_h = \frac{G_2(2)}{\log 2} = \frac{1}{\log 2}\int_0^\infty \frac{s}{Q(-2s)^b}\int_0^\infty e^{-zs}\tilde{g}(z)\,\mathrm{d}z\,\mathrm{d}s,$$

and

$$\varpi_h(t) = \frac{1}{\log 2}\sum_{k\in\mathbb{Z}\setminus\{0\}}\frac{G_2(2+\chi_k)}{\Gamma(2+\chi_k)}e^{2k\pi i t}.$$

By straightforward truncations, expansions and approximations, we obtain the following numerical values for $b = 1, \dots, 5$.

b	1	2	3	4	5
C_h	0.26600	0.13260	0.09004	0.06958	0.05781

More powerful means are needed to be developed if more degree of precision is required.

Generating functions. From (37), it follows that the moment generating function $M_n(y) := \mathbb{E}(e^{X_n y})$ can be recursively computed by the relation

$$M_{n+b}(y) = \frac{e^{ny}}{2^n} \sum_{0\le j\le n} \binom{n}{j} M_j(y) M_{n-j}(y) \qquad (n \ge 0),$$

with $M_n(y) = 1$ for $0 \le n < b$. The bivariate exponential generating function $F(z, y)$ then satisfies the equation

$$\frac{\partial^b}{\partial z^b} F(z, y) = F\left(\frac{e^y z}{2}, y\right)^2,$$

with $F^{(j)}(0, y) = 1$ for $0 \le j < b$, and we have the nonlinear equation for the Poisson generating function $\tilde{F}(z, y) := e^{-z} F(z, y)$

$$\sum_{0\le j\le b} \binom{b}{j} \tilde{F}^{(j)}(z, y) = e^{(e^y - 1)z} \tilde{F}\left(\frac{e^y z}{2}, y\right)^2, \tag{41}$$

with $\tilde{F}(0, y) = 1$.

From this form, the asymptotic analysis of the mean value and that of the variance proceed along exactly the same line we developed in the previous section. Thus we briefly sketch the principal steps of the analysis, leaving the details to the interested reader.

The expected value of X_n. From (41), we derive the following differential-functional equation for the Poisson generating function of the mean

$$\sum_{0\le j\le b} \binom{b}{j} \tilde{f}_1^{(j)}(z) = 2\tilde{f}_1(z/2) + z,$$

with the initial conditions $\tilde{f}_1^{(j)}(0) = 0$ for $0 \le j < b$.

Before applying the Laplace-Mellin approach, we need first a transfer-type result similar to Proposition 2.4.

Proposition 3.2 *Let $\tilde{f}(z)$ and $\tilde{g}(z)$ be entire functions satisfying*

$$\sum_{0\le j\le b} \binom{b}{j} \tilde{f}^{(j)}(z) = 2\tilde{f}(z/2) + \tilde{g}(z), \tag{42}$$

with $f(0) = 0$. Then

$$\tilde{g} \in \mathscr{JS} \iff \tilde{f} \in \mathscr{JS}.$$

Proof: (Sketch) The same proof as that for Proposition 2.4 applies *mutatis mutandis* to (42). The only difference is that we now have

$$f^{(b)}(z) = 2e^{z/2}f(z/2) + g(z),$$

where $f(z) := e^z\tilde{f}(z)$ and $g(z) := e^z\tilde{g}(z)$, so that (14) has the extended representation

$$\begin{aligned} f(z) &= \frac{1}{(b-1)!}\int_0^z (z-t)^{b-1}\left(2e^{t/2}f(t/2) + g(t)\right)\,\mathrm{d}t \\ &= \frac{z^b}{(b-1)!}\int_0^1 (1-t)^{b-1}\left(2e^{tz/2}f(tz/2) + g(tz)\right)\,\mathrm{d}t, \end{aligned}$$

and

$$\tilde{f}(z) = \frac{z^b}{(b-1)!}\int_0^1 (1-t)^b e^{-(1-t)z}\left(2\tilde{f}(tz/2) + \tilde{g}(z)\right)\,\mathrm{d}t.$$

All required estimates can be derived by the same arguments used there. $\square$

The Laplace transform of $\tilde{f}_1$ now satisfies the functional equation

$$(s+1)^b\mathscr{L}[\tilde{f}_1; s] = 4\mathscr{L}[\tilde{f}_1; 2s] + s^{-2},$$

for $\Re(s) > 0$. From this equation, we obtain

$$\mathscr{L}[\tilde{f}_1; s] = \frac{1}{s^2}\sum_{j\ge 0}\frac{1}{(s+1)^b\cdots(1+2^js)^b},$$

which extends (22). From this series and partial fraction expansions, we can derive a close-form expression for $\tilde{f}_1(z)$, which becomes messy especially for large b. Define as before $\bar{\mathscr{L}}[\tilde{f}_1; s] := \mathscr{L}[\tilde{f}_1; s]/Q(-s)^b$. Then we obtain

$$\bar{\mathscr{L}}[\tilde{f}_1; s] = 4\bar{\mathscr{L}}[\tilde{f}_1; 2s] + \frac{1}{Q(-2s)^b s^2}.$$

This relation is almost the same as (26). Thus the same Mellin analysis given there carries over and we deduce that

$$\begin{aligned} \mathscr{L}[\tilde{f}_1; s] &= \frac{1}{s^2}\log_2\frac{1}{s} + \frac{1}{s^2}\left(\frac{1}{2} + \frac{c_4}{\log 2} + \frac{1}{\log 2}\sum_{k\in\mathbb{Z}\setminus\{0\}} G_1(2+\chi_k)s^{-\chi_k}\right) \\ &\quad + \frac{b}{s}\log\frac{1}{s} + O(|s|^{-1}), \end{aligned}$$

uniformly for $|s| \to 0$ and $|\arg(s)| \le \pi - \varepsilon$, where

$$G_1(\omega) := \int_0^\infty \frac{s^{\omega-3}}{Q(-2s)^b}\,\mathrm{d}s,$$

and

$$
\begin{aligned}
c_4 &:= \lim_{\omega\to 2}\left(G_1(\omega) - \frac{1}{\omega-2}\right)\\
&= \int_0^1 \frac{1}{s}\left(\frac{1}{Q(-2s)^b} - 1\right) \mathrm{d}s + \int_1^\infty \frac{1}{sQ(-2s)^b}\,\mathrm{d}s.
\end{aligned}
$$

Consequently, by the Laplace inversion,

$$
\tilde f_1(z) = (z+b)\log_2 z + z\left(\frac{1}{2} + \frac{G_1(2)+\gamma-1}{\log 2} + \frac{1}{\log 2}\sum_{k\in\mathbb{Z}\setminus\{0\}} \frac{G_1(2+\chi_k)}{\Gamma(2+\chi_k)} z^{\chi_k}\right) + O(1), \quad (43)
$$

uniformly for $|z|\to\infty$ and $|\arg(z)| \le \pi/2-\varepsilon$. From this and Propositions 3.2 and 2.2, we obtain

$$
\mathbb{E}(X_n) = \sum_{0\le j<2k} \frac{\tilde f^{(j)}(n)}{j!}\,\tau_j(n) + O\left(n^{-1+k}\right),
$$

for any $k = 1, 2, \dots$. Finally,

$$
\mathbb{E}(X_n) = (n+b)\log_2 n + n\left(\frac{1}{2} + \frac{G_1(2)+\gamma-1}{\log 2} + \frac{1}{\log 2}\sum_{k\in\mathbb{Z}\setminus\{0\}} \frac{G_1(2+\chi_k)}{\Gamma(2+\chi_k)} n^{\chi_k}\right) + O(1).
$$

Variance of X_n. The analysis here is again similar to that for the mean. Let $\tilde f_2(z)$ denote the Poisson generating function of the second moment $\mathbb{E}(X_n^2)$. Then, by (41),

$$
\sum_{0\le j\le b}\binom{b}{j}\tilde f_2^{(j)}(z) = 2\tilde f_2(z/2) + 2\tilde f_1(z/2)^2 + 4z\tilde f_1(z/2) + 2z\tilde f_1'(z/2) + z + z^2,
$$

with the first b Taylor coefficients zero. Define again

$$
\tilde V(z) = \tilde f_2(z) - \tilde f_1(z)^2 - z\tilde f_1'(z)^2.
$$

Then $\tilde V(z)$ satisfies

$$
\sum_{0\le j\le b}\binom{b}{j}\tilde V^{(j)}(z) = 2\tilde V(z/2) + \tilde g(z),
$$

where $\tilde g(z)$ is given in (38).

By the representations (39) and (43), we have

$$
\tilde g(z) = \begin{cases} O(|z|), & \text{as } |z|\to 0;\\ O(|z|^{-1}), & \text{as } |z|\to\infty, \end{cases}
$$

uniformly in the sector $|\arg(z)| \le \pi/2-\varepsilon$. This is similar to the corresponding estimate (34) in the analysis of the variance in the previous section. The same procedure there applies and we deduce (40).

3.2 Node-wise path-length (NPL)

We consider in this section the total node-wise path-length (NPL). Under the same independent Bernoulli model, we still use X_n to denote the NPL in a random b-DST of n binary strings with node capacity $b \geq 2$. Also let N_n stand for the total number of nodes (space requirement) in random b-DST of n strings. Despite its being one of the most natural shape measures for b-DSTs, the consideration of X_n here seems to be new. For N_n, it is known that the distribution is asymptotically normal with the mean and the variance both asymptotically n times a different smooth periodic function; see [31]. In contrast to (40) for the variance of KPL, what is unexpected and surprising here is that the variance of X_n is of order $n(\log n)^2$.

Theorem 3.3 *Assume $b \geq 2$. The mean of N_n and that of X_n satisfy the following asymptotic relations.*

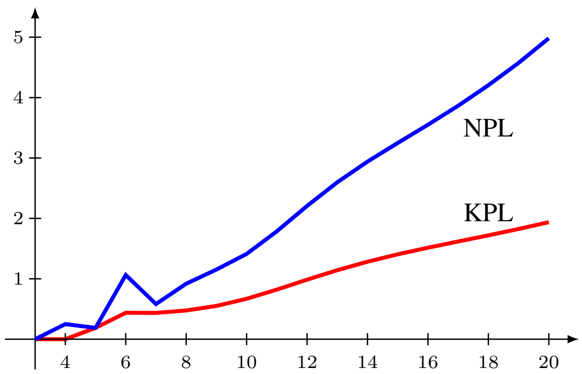
$$\begin{cases} \mathbb{E}(N_n) = nP_{1,0}(\log_2 n) + O(1), \\ \mathbb{E}(X_n) = n(\log_2 n)P_{1,0}(\log_2 n) + nP_{0,1}^{[2]}(\log_2 n) + (\log_2 n)P_{0,1}^{[3]}(\log_2 n) + O(1); \end{cases} \tag{44}$$

and the variances of N_n and X_n satisfy

$$\begin{cases} \mathbb{V}(N_n) = nP_{2,0}(\log_2 n) + O(1), \\ \mathrm{Cov}(N_n, X_n) = n(\log_2 n)P_{2,0}(\log_2 n) + nP_{1,1}^{[2]}(\log_2 n) + (\log n)P_{1,1}^{[3]}(\log_2 n) + O(1), \\ \mathbb{V}(X_n) = n(\log_2 n)^2 P_{2,0}(\log_2 n) + n(\log_2 n)P_{0,2}^{[2]}(\log_2 n) + nP_{0,2}^{[3]}(\log_2 n) \\ \qquad + (\log n)^2 P_{0,2}^{[4]}(\log_2 n) + (\log_2 n)P_{0,2}^{[5]}(\log_2 n) + O(1), \end{cases} \tag{45}$$

where the $P_{\cdot,\cdot}$'s are all computable, smooth, 1-periodic functions.

Intuitively, that the variance of NPL is larger than that of KPL can be seen from the definition of NPL, which depends on the random variable N_n (see (46)), while on the other hand, KPL depends on n only (in addition to on the two subtrees). The following figure shows the first few values of the variance of NPL and that of KPL.



We see that the variance of NPL increases faster than that of KPL.

Note that the periodic functions of the dominant terms are all equal, implying that the correlation coefficient of N_n and X_n is asymptotically 1.

On the other hand, the mean value $c_{1,0}$ of $P_{1,0}(t)$ is given by

$$c_{1,0} = \frac{1}{\log 2}\int_0^\infty \frac{(s+1)^{b-1}}{Q(-2s)^b}\,\mathrm{d}s;$$

numerical approximations to $c_{1,0}$ for the first few b are given as follows.

b	1	2	3	4	5	6
$c_{1,0}$	1	0.57470	0.40698	0.31594	0.25849	0.21885

Note that when $b = 1$

$$c_{1,0} = \frac{1}{\log 2}\int_0^\infty \frac{\mathrm{d}s}{Q(-2s)} = 1,$$

by (28), which is consistent with the fact that $N_n \equiv n$ in this case.

When $b = 2$, we see that about 42.5% of nodes on average contain two keys and 14% of nodes a single key. The storage utilization is thus not very bad.

From (44) and these numerical values, we see that, in contrast to the expected KPL, which is asymptotic to $n\log_2 n$ for all b, the expected NPL provides a better indication of the "shape variation" of random b-DSTs.

Our analysis is based on the following straightforward distributional recurrences

$$\begin{cases} N_{n+b} \stackrel{d}{=} N_{B_n} + N^*_{n-B_n} + 1, \\ X_{n+b} \stackrel{d}{=} X_{B_n} + X^*_{n-B_n} + N_{B_n} + N^*_{n-B_n}, \end{cases} \qquad (n \ge 0), \tag{46}$$

with the initial conditions $N_0 = 0, N_1 = \cdots = N_{b-1} = 1$ and $X_0 = \cdots = X_{b-1} = 0$. Here again $B_n \sim \text{Binomial}(n, 1/2), N_n \stackrel{d}{=} N_n^*, X_n \stackrel{d}{=} X_n^*$ and X_n, X_n^*, B_n as well as N_n, N_n^*, B_n are independent.

Generating functions. Define $M_n(u,v) = \mathbb{E}(e^{N_n u + X_n v})$. Then (46) translates into the recurrence

$$M_{n+b}(u,v) = e^u 2^{-n}\sum_{j=0}^n \binom{n}{j} M_j(u+v,v)M_{n-j}(u+v,v), \qquad (n \ge 0),$$

with $M_0(u,v) = 1, M_1(u,v) = \cdots = M_{b-1}(u,v) = e^u$. Next, let

$$F(z,u,v) := \sum_{n\ge 0} \frac{M_n(u,v)}{n!} z^n.$$

Then the recurrence relation gives

$$\frac{\partial^b}{\partial z^b} F(z,u,v) = e^u F\left(\frac{z}{2}, u+v, v\right)^2,$$

and the Poisson generating function $\tilde{F}(z,u,v) := e^{-z}F(z,u,v)$ satisfies

$$\sum_{0\le j\le b} \binom{b}{j} \frac{\partial^j}{\partial z^j}\tilde{F}(z,u,v) = e^u \tilde{F}\left(\frac{z}{2}, u+v, v\right)^2, \tag{47}$$

with the initial conditions $\tilde{F}(z,u,v) = 1 + (e^u - 1)\sum_{1\le j<b}(-1)^{j-1}z^j/j! + \cdots$.

For the moments, if we expand $\tilde{F}(z,u,v)$ in terms of u and v,

$$\tilde{F}(z,u,v) = \sum_{m\ge 0}\frac{1}{m!}\sum_{0\le j\le m}\binom{m}{j}\tilde{f}_{j,m-j}(z)u^j v^{m-j},$$

then $\tilde{f}_{j,m-j}(z)$ is the Poisson generating function of $\mathbb{E}(N_n^j X_n^{m-j})$. Thus all moments of X_n and N_n or their products can be computed by taking suitable derivatives of (47) with respect to u and v and then substituting $u = v = 0$.

Expected number of nodes and expected node-wise path length. By taking first derivatives of (47), we obtain

$$\begin{cases} \displaystyle\sum_{0\le j\le b}\binom{b}{j}\tilde{f}_{1,0}^{(j)}(z) = 2\tilde{f}_{1,0}(z/2) + 1, \\ \displaystyle\sum_{0\le j\le b}\binom{b}{j}\tilde{f}_{0,1}^{(j)}(z) = 2\tilde{f}_{0,1}(z/2) + 2\tilde{f}_{1,0}(z/2), \end{cases} \tag{48}$$

the initial conditions being $\tilde{f}_{1,0}(0) = 0$, $\tilde{f}_{1,0}^{(j)}(0) = (-1)^{j-1}$ for $1 \le j < b$ and $\tilde{f}_{0,1}^{(j)}(0) = 0$ for $0 \le j < b$.

We can apply the Laplace-Mellin approach as before, starting from the mean of N_n. Note that

$$\mathscr{L}[\tilde{f}^{(j)}; s] = s^j\mathscr{L}[\tilde{f}; s] - \sum_{0\le \ell<j} s^\ell \tilde{f}^{(j-1-\ell)}(0) \qquad (j = 0, 1, \dots),$$

provided that the Laplace transform exists for $\Re(s) > 0$. This gives

$$(s+1)^b\mathscr{L}[\tilde{f}_{1,0}; s] = 4\mathscr{L}[\tilde{f}_{1,0}; 2s] + \tilde{g}_{1,0}^\star(s),$$

where

$$\begin{aligned} \tilde{g}_{1,0}^\star(s) &:= \frac{1}{s} + \sum_{0\le\ell\le b-2} s^\ell \sum_{\ell\le j\le b-2}\binom{b}{j+2}\tilde{f}_{1,0}^{(j+1-\ell)}(0) \\ &= \frac{1}{s} + \sum_{1\le j<b}\binom{b-1}{j}s^{j-1} \\ &= s^{-1}(s+1)^{b-1}. \end{aligned}$$

Unlike all previous cases, iterating this functional equation leads to a divergent series. Although this problem can be solved by subtracting a sufficient number of initial terms of $\tilde{f}_{1,0}(z)$, the approach we use does not rely on this and avoids completely such a consideration.

Let $\bar{\mathscr{L}}[\tilde{f}_{1,0}; s] := \mathscr{L}[\tilde{f}_{1,0}; s]/Q(-s)^b$. Then

$$\mathscr{M}[\bar{\mathscr{L}}[\tilde{f}_{1,0}; s]; \omega] = \frac{G_{1,0}(\omega)}{1 - 2^{2-\omega}}, \qquad (\Re(\omega) > 2),$$

where

$$G_{1,0}(\omega) := \int_0^\infty \frac{s^{\omega-2}}{Q(-2s)^b}(s+1)^{b-1}\,\mathrm{d}s,$$

for $\Re(\omega) > 1$.

From this, we deduce that

$$\tilde{f}_{1,0}(z) = zP_{1,0}(\log_2 z) + O(1), \tag{49}$$

uniformly for $|z| \to \infty$ and $|\arg(z)| \le \pi/2 - \varepsilon$, where $P_{1,0}(t)$ is a periodic function with the Fourier series representation

$$P_{1,0}(t) := \frac{1}{\log 2}\sum_{k\in\mathbb{Z}} \frac{G_{1,0}(2+\chi_k)}{\Gamma(2+\chi_k)} e^{2k\pi i t},$$

the series being absolutely convergent. From this we deduce the first approximation of (44).

We now turn to the expected NPL $\mathbb{E}(X_n)$. By (48), we have

$$(s+1)^b \mathscr{L}[\tilde{f}_{0,1}; s] = 4\mathscr{L}[\tilde{f}_{0,1}; 2s] + 4\mathscr{L}[\tilde{f}_{1,0}; 2s].$$

Let $\bar{\mathscr{L}}[\tilde{f}_{0,1}; s] := \mathscr{L}[\tilde{f}_{0,1}; s]/Q(-s)^b$. Then

$$\mathscr{M}[\bar{f}_{0,1}; \omega] = \frac{2^{2-\omega}G_{1,0}(\omega)}{(1-2^{2-\omega})^2}, \qquad (\Re(\omega) > 2).$$

From this we deduce that

$$\tilde{f}_{0,1}(z) = z(\log_2 z)P_{0,1}^{[1]}(\log_2 z) + zP_{0,1}^{[2]}(\log_2 z) + (\log_2 z)P_{0,1}^{[4]}(\log_2 z) + O(1), \tag{50}$$

uniformly for $|z| \to \infty$ and $|\arg(z)| \le \pi/2 - \varepsilon$, where $P_{0,1}^{[1]}(t), P_{0,1}^{[2]}(t), P_{0,1}^{[4]}(t)$ are smooth, 1-periodic functions whose Fourier coefficients are given by

$$\begin{aligned}
P_{0,1}^{[1]}(t) &= P_{1,0}(t) = \frac{1}{\log 2}\sum_{k\in\mathbb{Z}} \frac{G_{1,0}(2+\chi_k)}{\Gamma(2+\chi_k)}\, e^{2k\pi i t},\\
P_{0,1}^{[2]}(t) &= -\frac{1}{(\log 2)^2}\sum_{k\in\mathbb{Z}} \frac{G_{1,0}'(2+\chi_k)\psi(2+\chi_k) - G_{1,0}(2+\chi_k)}{\Gamma(2+\chi_k)}\, e^{2k\pi i t},\\
P_{0,1}^{[4]}(t) &= \frac{b}{\log 2}\sum_{k\in\mathbb{Z}} \frac{G_{1,0}(2+\chi_k)}{\Gamma(1+\chi_k)}\, e^{2k\pi i t}.
\end{aligned}$$

Here $\psi(z)$ denotes the derivative of $\log\Gamma(z)$ and all series are absolutely convergent. This proves (44).

Variance. Taking second derivatives in (47) and substituting $u = v = 0$ gives

$$\begin{cases} \sum_{0\le j\le b}\binom{b}{j}\tilde f_{2,0}^{(j)}(z) = 2\tilde f_{2,0}(z/2) + 2\tilde f_{1,0}(z/2)^2 + 4\tilde f_{1,0}(z/2) + 1, \\ \sum_{0\le j\le b}\binom{b}{j}\tilde f_{1,1}^{(j)}(z) = 2\tilde f_{1,1}(z/2) + 2\tilde f_{2,0}(z/2) + 2(\tilde f_{1,0}(z/2) + \tilde f_{0,1}(z/2))(\tilde f_{1,0}(z/2) + 1), \\ \sum_{0\le j\le b}\binom{b}{j}\tilde f_{0,2}^{(j)}(z) = 2\tilde f_{0,2}(z/2) + 4\tilde f_{1,1}(z/2) + 2\tilde f_{0,2}(z/2) + 2(\tilde f_{1,0}(z/2) + \tilde f_{0,1}(z/2))^2, \end{cases}$$

with the initial conditions $\tilde f_{2,0}^{(j)}(0) = (-1)^{j-1}$ for $1 \le j < b$ and $\tilde f_{2,0}(0) = \tilde f_{1,1}^{(j)}(0) = \tilde f_{0,2}^{(j)}(0) = 0$, for $0 \le j < b$.

The remaining calculations follow the same pattern of proof we used above but become much more involved. We begin with

$$\begin{cases} \tilde V(z) = \tilde f_{2,0}(z) - \tilde f_{1,0}(z)^2 - z\tilde f_{1,0}'(z)^2, \\ \tilde U(z) = \tilde f_{1,1}(z) - \tilde f_{1,0}(z)\tilde f_{0,1}(z) - z\tilde f_{1,0}'(z)\tilde f_{0,1}'(z), \\ \tilde W(z) = \tilde f_{0,2}(z) - \tilde f_{0,1}(z)^2 - z\tilde f_{0,1}'(z)^2. \end{cases}$$

Then we deduce

$$\begin{cases} \sum_{0\le j\le b}\binom{b}{j}\tilde V^{(j)}(z) = 2\tilde V(z/2) + \tilde g_{2,0}(z), \\ \sum_{0\le j\le b}\binom{b}{j}\tilde U^{(j)}(z) = 2\tilde U(z/2) + \tilde g_{1,1}(z), \\ \sum_{0\le j\le b}\binom{b}{j}\tilde W^{(j)}(z) = 2\tilde W(z/2) + \tilde g_{0,2}(z), \end{cases}$$

where

$$\begin{cases} \tilde g_{2,0}(z) = \left(\sum_{0\le j\le b}\binom{b}{j}\tilde f_{1,0}^{(j)}(z)\right)^2 + z\left(\sum_{0\le j\le b}\binom{b}{j}\tilde f_{1,0}^{(j+1)}(z)\right)^2 \\ \qquad - \sum_{0\le j\le b}\binom{b}{j}\left(\tilde f_{1,0}(z)^2 + z\tilde f_{1,0}'(z)^2\right)^{(j)}, \\ \tilde g_{1,1}(z) = 2\tilde V(z/2) + \left(\sum_{0\le j\le b}\binom{b}{j}\tilde f_{1,0}^{(j)}(z)\right)\left(\sum_{0\le j\le b}\binom{b}{j}\tilde f_{0,1}^{(j)}(z)\right) \\ \qquad + z\left(\sum_{0\le j\le b}\binom{b}{j}\tilde f_{1,0}^{(j+1)}(z)\right)\left(\sum_{0\le j\le b}\binom{b}{j}\tilde f_{0,1}^{(j+1)}(z)\right) \\ \qquad - \sum_{0\le j\le b}\binom{b}{j}\left(\tilde f_{1,0}(z)\tilde f_{0,1}(z) + z\tilde f_{1,0}'(z)\tilde f_{0,1}'(z)\right)^{(j)}, \\ \tilde g_{0,2}(z) = 4\tilde U(z/2) + 2\tilde V(z/2) + \left(\sum_{0\le j\le b}\binom{b}{j}\tilde f_{0,1}^{(j)}(z)\right)^2 \\ \qquad + z\left(\sum_{0\le j\le b}\binom{b}{j}\tilde f_{0,1}^{(j+1)}(z)\right)^2 - \sum_{0\le j\le b}\binom{b}{j}\left(\tilde f_{0,1}(z)^2 + z\tilde f_{0,1}'(z)^2\right)^{(j)}. \end{cases}$$

The initial conditions are $\tilde V(0) = \tilde U^{(j)}(0) = \tilde W^{(j)}(0) = 0$ for $0 \le j < b$ and

$$\tilde V^{(j)}(0) = (-1)^j\left(1 + (j-2)2^{j-1}\right), \qquad (1 \le j \le b).$$

From (49), (50) and Ritt's theorem (see [54]), we have

$$\begin{cases} \tilde g_{2,0}(z) = O\left(|z|^{-1}\right), \\ \tilde g_{1,1}(z) - 2\tilde V(z/2) = O\left(|z|^{-1}\right), \\ \tilde g_{0,2}(z) - 4\tilde U(z/2) - 2\tilde V(z/2) = O\left(|z|^{-1}\right), \end{cases}$$

uniformly for $|z| \to \infty$ and $|\arg(z)| \le \pi/2 - \varepsilon$. Let $\bar{\mathscr{L}}[\tilde A; s] := \mathscr{L}[\tilde A; s]/Q(-s)^b$, where $\tilde A \in \{\tilde V, \tilde U, \tilde W\}$. Then we obtain, for $\Re(\omega) > 2$,

$$\begin{cases} \mathscr{M}[\bar{\mathscr{L}}[\tilde V; s]; \omega] = \dfrac{G_{2,0}(\omega)}{1 - 2^{2-\omega}}, \\ \mathscr{M}[\bar{\mathscr{L}}[\tilde U; s]; \omega] = \dfrac{2^{2-\omega}G_{2,0}(\omega)}{(1 - 2^{2-\omega})^2} + \dfrac{G_{1,1}(\omega)}{1 - 2^{2-\omega}}, \\ \mathscr{M}[\bar{\mathscr{L}}[\tilde W; s]; \omega] = \dfrac{2^{5-2\omega}G_{2,0}(\omega)}{(1 - 2^{2-\omega})^3} + \dfrac{2^{2-\omega}\left(2G_{1,1}(\omega) + G_{2,0}(\omega)\right)}{(1 - 2^{2-\omega})^2} + \dfrac{G_{1,1}(\omega) + G_{0,2}(\omega)}{1 - 2^{2-\omega}}, \end{cases}$$

where

$$\begin{cases} G_{2,0}(\omega) := \displaystyle\int_0^\infty \frac{s^{\omega-1}}{Q(-2s)^b} \left(\mathscr{L}[\tilde{g}_{2,0}; s] + \frac{(s+1)^{b-1} - (-1)^b\,(2b-3+(b-1)s)}{(s+2)^2} \right) \mathrm{d}s, \\ G_{1,1}(\omega) := \displaystyle\int_0^\infty \frac{s^{\omega-1}}{Q(-2s)^b} \int_0^\infty e^{-sz} \left(\tilde{g}_{1,1}(z) - 2\tilde{V}(z/2) \right) \mathrm{d}z\,\mathrm{d}s, \\ G_{0,2}(\omega) := \displaystyle\int_0^\infty \frac{s^{\omega-1}}{Q(-2s)^b} \int_0^\infty e^{-sz} \left(\tilde{g}_{0,2}(z) - 2\tilde{V}(z/2) - 4\tilde{U}(z/2) \right) \mathrm{d}z\,\mathrm{d}s, \end{cases}$$

with all functions analytic for $\Re(\omega) > 0$. Consequently, we deduce (45).

4 Digital search trees. II. More shape parameters.

We consider in this section four additional examples on DSTs whose variances are essentially linear. The same tools we use readily apply to b-DSTs, but we focus on DSTs because the results are easier to state and the asymptotic behaviors do not differ in essence with those for the more general b-DSTs the corresponding expressions of which are however much messier.

The first parameter we consider is the so-called w-parameter (see [16]), which is the sum of the subtree-size of the parent-node of each leaf (over all leaves)[(iii)]. Instead of w-parameter, we call it the *total peripheral path-length* (PPL), since it measures to some extent the fringe ampleness of the trees. Also this is in consistency with the two previous notions of path-length we distinguished.

Then we consider the number of leaves, which has previously been studied in details in [26, 31, 39] and which is well connected to PPL. Our expression for the variance simplifies known ones.

Yet another notion of path-length we consider here is the so-called *Colless index* in phylogenetics, which is the sum of the absolute difference of the two subtree-sizes of each node (over all nodes). We call this index the *total differential path-length* (DPL) as it clearly indicates the balance or symmetry of the tree. Another widely used measure of imbalance in phylogenetics is the *Sackin index*, which is nothing but the external path-length.

The last example we consider is the *weighted path-length* (WPL), which often arises in coding, optimization and many related problems.

The orders of the means and the variances exhibited by all the shape parameters we study in this paper are listed in Table 1.

4.1 Peripheral path-length (PPL)

The PPL (or w-parameter) was introduced in [16], the motivations arising from the analysis of compression algorithms. We start from the *fringe-size* of a leaf node λ, which is defined to be the size of the subtree rooted at its parent-node; see Figure 9. The PPL of a tree is then defined to be the sum of the fringe-sizes of all leaf-nodes. Let X_n denote the PPL in a DST built from n random binary strings under our usual independent Bernoulli model.

Drmota et al. showed in [16] that

$$\mathbb{E}(X_n) = n\left(C_w + \varpi_w(\log_2 n)\right) + o(n), \tag{51}$$

[(iii)] The *leaves* or *leaf-nodes* of a tree are nodes without any descendants.

Fig. 9: *The two possible configurations of the fringe of a leaf: the fringe-size (or w-parameter) equals $|T|+2$. Note that T may be empty.*

where

$$C_w := \sum_{\ell\ge0}\frac{(\ell+1)(\ell-2)}{Q_\ell 2^\ell}\left(\sum_{k\ge1}\frac{1}{2^{\ell+k}-1}-1\right)+\frac{1}{\log 2}\sum_{\ell\ge0}\frac{2\ell-1}{Q_\ell 2^\ell}.$$

Note that by (24), we have the identities

$$\sum_{\ell\ge0}\frac{(\ell+1)(\ell-2)}{Q_\ell 2^\ell}=\frac{1}{Q_\infty}\left(\sum_{j\ge1}\frac{1}{(2^j-1)^2}+\left(\sum_{j\ge1}\frac{1}{2^j+1}\right)^2-2\right),$$
$$\sum_{\ell\ge0}\frac{2\ell-1}{Q_\ell 2^\ell}=\frac{1}{Q_\infty}\left(\sum_{j\ge1}\frac{2}{2^j-1}-1\right).$$

The asymptotic behavior (51) is to be compared with the $n\log n$-order exhibited by most other log-trees such as binary search trees and recursive trees; see [16]. It reflects that most fringes of random DSTs are small in size; see Figure 3. Indeed, since the expected number of leaves is also asymptotic to n times a periodic function, the result (51) implies that the average size of a fringe in random DSTs is bounded. We show that the standard deviation is also small.

Define

$$\begin{aligned}\tilde g_2(z) &:= z\tilde f_1''(z)^2-\frac{z}{16}\,e^{-z}\left(z^4+4z^3+16z^2-8z+64\right)\\ &\quad-\frac{z}{4}\,e^{-z/2}\Big(4(z+4)\tilde f_1(z/2)-2(z^2+2z+8)\tilde f_1'(z/2)-(z+2)(z+8)\Big),\end{aligned}\tag{52}$$

where $\tilde f_1(z)$ represents as usual the Poisson generating function of $\mathbb{E}(X_n)$. Let $G_2(\omega)$ denote the Mellin transform of $\mathscr{L}[\tilde g_2;s]/Q(-2s)$.

Theorem 4.1 *The mean and the variance of the total PPL X_n of random DSTs of n strings satisfy*

$$\begin{aligned}\mathbb{E}(X_n)&=n\left(C_w+\varpi_w(\log_2 n)\right)+O(1),\\ \mathbb{V}(X_n)&=nP_w(\log_2 n)+O(1),\end{aligned}\tag{53}$$

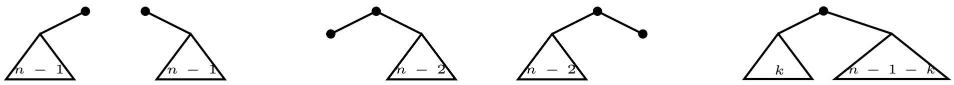
where $P_w(t)$ is a smooth, 1-periodic function with the Fourier series expansion

$$P_w(t)=\frac{1}{\log 2}\sum_{k\in\mathbb{Z}}\frac{G_2(2+\chi_k)}{\Gamma(2+\chi_k)}\,e^{2k\pi it},$$

the series being absolutely convergent.

We provide only the major steps of the proof since it follows the same approach we developed above.

Recurrence and generating functions. By definition and by conditioning on the size of one of the subtrees of the root, we have the following different configurations



from which we derive the recurrence for the PPL

$$X_n \stackrel{d}{=} \begin{cases} X_{n-1}, & \text{with probability } 2^{2-n}; \\ n + X_{n-2}, & \text{with probability } (n-1)2^{2-n}; \\ X_k + X^*_{n-1-k}, & \text{with probability } 2^{1-n}\binom{n-1}{k},\ 2 \le k \le n-3, \end{cases}$$

where $X_0 = X_1 = 0, X_2 = 2$ and X_3 has the distribution

$$X_3 = \begin{cases} 6, & \text{with probability } 1/2; \\ 2, & \text{with probability } 1/2. \end{cases}$$

From this recurrence, it follows that the bivariate Poisson generating function

$$\tilde{F}(z, y) := e^{-z} \sum_{n \ge 0} \frac{\mathbb{E}(e^{X_n y})}{n!}\, z^n$$

satisfies the nonlinear equation

$$\begin{aligned} \tilde{F}(z, y) + \frac{\partial}{\partial z}\tilde{F}(z, y) = \tilde{F}\left(\frac{z}{2}, y\right)^2 + ze^{2y+e^y z/2-z}\tilde{F}\left(\frac{e^y z}{2}, y\right) \\ - ze^{-z/2}\tilde{F}\left(\frac{z}{2}, y\right) + \frac{z^2}{4}\, e^{-z}\left(e^{3y} - 1\right)^2, \end{aligned} \tag{54}$$

with the initial condition $\tilde{F}(0, y) = 1$.

The expected PPL. By (54), we obtain the differential-functional equation for $\tilde{f}_1(z)$ by taking derivative with respect to y and then substituting $y = 1$, giving

$$\tilde{f}_1(z) + \tilde{f}_1'(z) = 2\tilde{f}_1(z/2) + z(2 + z/2)e^{-z/2}, \tag{55}$$

with $f_1(0) = 0$. The Laplace transform of $\tilde{f}_1$ satisfies

$$\begin{aligned} \mathscr{L}[\tilde{f}_1; s] &= \frac{4}{s+1}\,\mathscr{L}[\tilde{f}_1; s] + \frac{16}{(1+2s)^3} \\ &= 16 \sum_{k \ge 0} \frac{4^k}{(s+1)\cdots(2^{k-1}s+1)(2^{k+1}s+1)^3}. \end{aligned}$$

Then a straightforward application of the Laplace-Mellin-de-Poissonization approach yields

$$\mathbb{E}(X_n) = \frac{n}{\log 2} \sum_{k\in\mathbb{Z}} \frac{G_1(2+\chi_k)}{\Gamma(2+\chi_k)}\, n^{\chi_k} + O(1),$$

where

$$G_1(\omega) := 16 \int_0^\infty \frac{s^{\omega-1}}{Q(-s)(2s+1)^3}\,\mathrm{d}s \qquad (\Re(\omega) > 0).$$

The $O(1)$-term can be further refined by the same analysis. In particular, we get an alternative expression for C_w

$$C_w = \frac{G_1(2)}{\log 2} = \frac{16}{\log 2}\int_0^\infty \frac{s}{Q(-s)(2s+1)^3}\,\mathrm{d}s \approx 1.10302\,66959\cdots.$$

That the two expressions of C_w are identical can be proved by standard calculus of residues; see [24] for similar details.

The variance of the PPL. Again from (54), we derive the equation for the Poisson generating function $\tilde f_2(z)$ of the second moment of X_n

$$\begin{aligned}\tilde f_2(z) + \tilde f_2'(z) &= 2\tilde f_2(z/2) + 2\tilde f_1(z/2)^2 + \frac{9}{2}\, z^2 e^{-z}\\ &\quad + z e^{-z/2}\left((z+4)\tilde f_1(z/2) + z\tilde f_1'(z/2) + \frac{z^2+10z+16}{4}\right),\end{aligned} \tag{56}$$

with $\tilde f_2(0) = 0$.

Let $\tilde V(z) = \tilde f_2(z) - \tilde f_1(z)^2 - z\tilde f_1'(z)^2$. Then, by (55), (56) and Lemma 2.7,

$$\tilde V(z) + \tilde V'(z) = 2\tilde V(z/2) + \tilde g_2(z),$$

with $\tilde V(0) = 0$, where $\tilde g_2$ is defined in (52).

Applying again the Laplace-Mellin-de-Poissonization approach, we deduce (53). In particular, the mean value of the periodic function P_w is given by

$$\frac{G_2(2)}{\log 2} = \frac{1}{\log 2}\int_0^\infty \frac{s}{Q(-2s)}\int_0^\infty e^{-zs}\tilde g_2(z)\,\mathrm{d}z\,\mathrm{d}s.$$

4.2 The number of leaves

The leaves of a tree are the locations where the nodes holding new-coming keys will be connected; thus different types of data fields can be used to save memory, notably for b-DSTs. The number of leaves then provides a quick and simpler look at the "fringes" of a tree. Such nodes are sometimes referred to as the external-internal nodes or internal endnodes in the literature; see [16, 26, 41, 56].

Let X_n denote the number of leaves in a random DST of n keys. Then X_n satisfies the recurrence

$$X_{n+1} \stackrel{d}{=} X_{B_n} + X^*_{n-B_n} \qquad (n \ge 1), \tag{57}$$

with $X_0 = 0$ and $X_1 = 1$, where $B_n \sim \text{Binomial}(n; 1/2)$.

Flajolet and Sedgewick [26], solving an open question raised by Knuth, showed that

$$\mathbb{E}(X_n) = n\left(C_{fs} + \varpi_{fs}(\log_2 n)\right) + O(n^{1/2}),$$

where $\varpi_{fs}(t)$ is a smooth, 1-periodic function and

$$\begin{aligned} C_{fs} &= 1 + \sum_{k\ge1} \frac{k}{Q_k 2^k} \sum_{1\le j\le k} \frac{1}{2^j-1} - \frac{1}{Q_\infty}\left(\frac{1}{\log 2} + \left(\sum_{k\ge1}\frac{1}{2^k-1}\right)^2 - \sum_{k\ge1}\frac{1}{2^k-1}\right) \\ &\approx 0.37204\,86812\cdots. \end{aligned}$$

A finer approximation, together with the alternative (and numerically better) expression

$$C_{fs} = 1 + \sum_{k\ge1}\frac{1}{2^k-1} - \frac{1}{Q_\infty}\left(\frac{1}{\log 2} + \sum_{k\ge1}\frac{(-1)^k k}{Q_k(2^k-1)2^{k(k+1)/2}}\right),$$

was derived by Kirschenhofer and Prodinger [39]; see also [56]. They proved additionally the asymptotic linearity of the variance

$$\mathbb{V}(X_n) \sim n\left(C_{kp} + \varpi_{kp}(\log_2 n)\right),$$

where ϖ_{kp} is a smooth, 1-periodic function with mean zero and a long, complicated expression is given for the leading constant C_{kp}. We derive different forms for these two asymptotic approximations.

Define

$$\tilde{g}_2(z) = z\tilde{f}_1''(z)^2 + e^{-z}\Big(1 - e^{-z}(1+z) + 2z\tilde{f}_1'(z/2) - 4\tilde{f}_1(z/2)\Big), \tag{58}$$

where $\tilde{f}_1(z) := e^{-z}\sum_{n\ge0}\mathbb{E}(X_n)z^n/n!$.

Theorem 4.2 *The mean and the variance of the number of leaves are both asymptotically linear with the approximations*

$$\begin{aligned} \mathbb{E}(X_n) &= \frac{n}{\log 2}\sum_{k\in\mathbb{Z}} \frac{G_1(2+\chi_k)}{\Gamma(1+\chi_k)}\, n^{\chi_k} + O(1), \\ \mathbb{V}(X_n) &= \frac{n}{\log 2}\sum_{k\in\mathbb{Z}} \frac{G_2(2+\chi_k)}{\Gamma(2+\chi_k)}\, n^{\chi_k} + O(1), \end{aligned}$$

where the two series are absolutely convergent with G_1, G_2 defined by

$$\begin{aligned} G_1(\omega) &= \int_0^\infty \frac{s^{\omega-1}}{(s+1)Q(-2s)}\,\mathrm{d}s, \\ G_2(\omega) &= \int_0^\infty \frac{s^{\omega-1}}{Q(-2s)}\int_0^\infty e^{-zs}\tilde{g}_2(z)\,\mathrm{d}z\,\mathrm{d}s, \end{aligned}$$

for $\Re(\omega) > 0$.

We see in particular that

$$C_{fs} = \frac{1}{\log 2}\int_0^\infty \frac{s}{(s+1)Q(-2s)}\,\mathrm{d}s,$$
$$C_{kp} = \frac{1}{\log 2}\int_0^\infty \frac{s}{Q(-2s)}\int_0^\infty e^{-zs}\tilde{g}_2(z)\,\mathrm{d}z\,\mathrm{d}s. \tag{59}$$

Sketch of proof. From (57), we derive the equation for the bivariate generating function $\tilde{F}(z,y) := e^{-z}\sum_{n\ge 0}\mathbb{E}(e^{X_n y})z^n/n!$

$$\tilde{F}(z,y) + \frac{\partial}{\partial z}\tilde{F}(z,y) = \tilde{F}\left(\frac{z}{2},y\right)^2 + (e^y-1)\,e^{-z},$$

with $\tilde{F}(0,y)=1$. Then the Poisson generating functions of the first two moments satisfy

$$\tilde{f}_1(z) + \tilde{f}_1'(z) = 2\tilde{f}_1(z/2) + e^{-z}, \tag{60}$$
$$\tilde{f}_2(z) + \tilde{f}_2'(z) = 2\tilde{f}_2(z/2) + 2\tilde{f}_1(z/2)^2 + e^{-z},$$

with $\tilde{f}_1(0)=\tilde{f}_2(0)$. Consequently, the function $\tilde{V}(z) := \tilde{f}_2(z) - \tilde{f}_1(z)^2 - z\tilde{f}_1'(z)^2$ satisfies

$$\tilde{V}(z) + \tilde{V}'(z) = 2\tilde{V}(z/2) + \tilde{g}_2(z),$$

with $\tilde{V}(0)=0$, where $\tilde{g}_2$ is given in (58). The remaining analysis follows the same pattern as above and is omitted.

We provide instead some details for the numerical evaluation of the constant C_{kp} as defined in (59), which is similar to the case of internal path-length of DSTs.

By applying the Laplace transform to both sides of (60) and by iteration, we get

$$\mathscr{L}[\tilde{f}_1;s] = \sum_{k\ge 0}\frac{4^k}{(s+1)(2s+1)\cdots(2^{k-1}s+1)(2^k s+1)^2}.$$

Since the inverse Laplace transform derived from the partial fraction expansion of this series is divergent, we consider the function $\hat{f}_1(z) := \tilde{f}_1(z) - z + z^2/2$ for which the equation (60) becomes

$$\hat{f}_1(z) + \hat{f}_1'(z) = 2\hat{f}_1(z/2) - 1 + z + \frac{z^2}{4} + e^{-z},$$

with $\hat{f}_1(0)=0$, and we have

$$\mathscr{L}[\hat{f}_1;s] = \frac{1}{2s^3}\sum_{k\ge 0}\frac{3\cdot 2^k s+1}{2^k(s+1)\cdots(2^{k-1}s+1)(2^k s+1)^2}.$$

Then by the partial fraction expansion

$$\frac{3\cdot 2^k s+1}{(s+1)\cdots(2^{k-1}s+1)(2^k s+1)^2} = \sum_{0\le \ell<k}\frac{(-1)^{k-\ell}(3\cdot 2^{k-\ell}-1)2^{-\binom{k-\ell+1}{2}}}{(2^{k-\ell}-1)Q_\ell Q_{k-\ell}}\cdot\frac{1}{2^\ell s+1}$$
$$+\frac{1}{Q_k}\left(3+2\sum_{1\le j\le k}\frac{1}{2^j-1}\right)\frac{1}{2^k s+1} - \frac{2}{Q_k(2^k s+1)^2},$$

we obtain

$$\mathscr{L}[\hat{f}_1; s] = \frac{1}{2s^3} \sum_{\ell \ge 0} \frac{1}{2^\ell Q_\ell} \left(\frac{\delta_\ell}{2^\ell s + 1} - \frac{2}{(2^\ell s + 1)^2} \right),$$

where

$$\delta_\ell = 3 + 2 \sum_{1 \le j \le \ell} \frac{1}{2^j - 1} + \sum_{j \ge 1} \frac{(-1)^j (3 \cdot 2^j - 1) 2^{-\binom{j+1}{2}}}{(2^j - 1) 2^j Q_j}.$$

Obviously, $\lim_{\ell \to \infty} \delta_\ell = 4$. Now, by the inverse Laplace transform,

$$\begin{aligned}\hat{f}_1(z) = \frac{1}{2} \sum_{\ell \ge 0} \frac{1}{Q_\ell} \Bigg(& 2^\ell \delta_\ell \left(1 - \frac{z}{2^\ell} + \frac{z^2}{2^{2\ell+1}} - e^{-z/2^\ell} \right) \\ & - 2^{\ell+1} \left(3 - \frac{z}{2^{\ell-1}} + \frac{z^2}{2^{2\ell+1}} - 3e^{-z/2^\ell} \right) + 2ze^{-z/2^\ell} \Bigg),\end{aligned}$$

which converges for all z; also from [26] we have

$$\hat{f}_1(z) = \sum_{n \ge 3} \frac{(-1)^{n-1} z^n}{n!} Q(n-2) \sum_{0 \le j \le n-2} \frac{1}{Q(j)}.$$

Then the first and the second derivatives are given by

$$\begin{aligned}\hat{f}_1'(z) &= \frac{1}{2} \sum_{\ell \ge 0} \frac{1}{Q_\ell} \left(\delta_\ell \left(-1 + z/2^\ell + e^{-z/2^\ell} \right) + 4 - \frac{z}{2^{\ell-1}} - 4e^{-z/2^\ell} - \frac{z}{2^{\ell-1}} e^{-z/2^\ell} \right), \\ \hat{f}_1''(z) &= \frac{1}{2} \sum_{\ell \ge 0} \frac{1}{2^\ell Q_\ell} \left(\delta_\ell \left(1 - e^{-z/2^\ell} \right) - 2 + 2e^{-z/2^\ell} + \frac{z}{2^{\ell-1}} e^{-z/2^\ell} \right).\end{aligned}$$

Now the constant C_{kp} can be expressed in terms of the integrals of $\hat{f}_1$ as follows.

$$\begin{aligned}(\log 2) C_{kp} = & \int_0^\infty \frac{s}{Q(-2s)(s+1)(s+2)^2} \, \mathrm{d}s + \int_0^\infty \frac{s}{Q(-2s)} \int_0^\infty e^{-zs} z (\hat{f}_1''(z) - 1)^2 \, \mathrm{d}z \, \mathrm{d}s \\ & + 2 \int_0^\infty \frac{s}{Q(-2s)} \int_0^\infty e^{-z(s+1)} \left(z - \frac{1}{s+1} \right) (\hat{f}_1'(z/2) - z) \, \mathrm{d}z \, \mathrm{d}s.\end{aligned}$$

And we get $C_{kp} \approx 0.034203 \cdots$.

A general weighted sum of node-types for b-DSTs. For $b \ge 2$, we can consider $X_n^{[j]}, 1 \le j \le b$, the number of leaves containing j records in a random b-DST with bucket capacity b built from n records. Let also $X_n^{[b+1]}$ be the number of internal (non-leaf) nodes. Define

$$X_n = \sum_{1 \le j \le b+1} a_j X_n^{[j]},$$

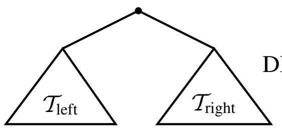
where $a_1, \dots, a_{b+1}$ are arbitrary real numbers. By a straightforward computation

$$\sum_{0\le j\le b} \binom{b}{j} \frac{\partial^j}{\partial z^j} \tilde F(z,y) = e^{a_{b+1}y} \tilde F\left(\frac{z}{2}, y\right)^2 + e^{-z}\left(e^{a_b y} - e^{a_{b+1}y}\right),$$

with $\tilde F(0,y) = 1$. Then our approach can be applied and leads to the same type of results as Theorem 4.2 with different G_1 and G_2; the resulting expressions for the variance are more explicit and simpler than those given in [31].

4.3 Colless index: the differential path-length (DPL)

The DPL of a tree is defined to be the sum over all nodes of the absolute difference of the two subtree-sizes of each node as depicted below.



$$\text{DPL} = \sum_{\text{all nodes}} |\mathcal{T}_{\text{left}} - \mathcal{T}_{\text{right}}|$$

Properties of such a path length in random binary search trees have long been investigated in the systematic biology literature; see [4] and the references therein.

Let X_n denote the DPL of a random DST of n input-strings. Then by definition and by our independence assumption, we have the recurrence for the moment generating function

$$M_{n+1}(y) = 2^{-n} \sum_{0\le j\le n} \binom{n}{j} M_j(y) M_{n-j}(y) e^{|n-2j|y} \qquad (n \ge 0), \tag{61}$$

with $M_0(y) = 1$.

Let also

$$\tilde g_2(z) := z\tilde f_1''(z)^2 + z - \tilde h_1(z)^2 - z\tilde h_1'(z)^2 - 4\tilde h_1(z)\tilde f(z/2) - 2z\tilde h_1'(z)\tilde f_1'(z/2) + 4\tilde h_c(z),$$

where $\tilde f_1(z)$ is the Poisson generating function of $\mathbb{E}(X_n)$ and $\tilde h_c(z)$ is defined by

$$\tilde h_c(z) := e^{-z} \sum_{n\ge 0} \frac{(z/2)^n}{n!} \sum_{0\le k\le n} \binom{n}{k} \mathbb{E}(X_k)|n-2k|. \tag{62}$$

Theorem 4.3 *The mean and the variance of the DPL of random DSTs satisfy the asymptotic relations*

$$\mathbb{E}(X_n) = nP_{d,\mu}(\log_2 n) - \frac{\sqrt{2n}}{\sqrt{\pi}(\sqrt{2}-1)} + O(1), \tag{63}$$

$$\mathbb{V}(X_n) = \left(1 - \frac{2}{\pi}\right) n\log_2 n + nP_{d,\sigma}(\log_2 n) + O(n^{1/2}), \tag{64}$$

where $P_{d,\mu}$ and $P_{d,\sigma}$ are explicitly computable, smooth, 1-periodic functions.

These results are to be compared with the known results for random binary search trees for which the DPL has mean of order $n\log n$ and variance of order n^2; see [4].

Expected DPL. The approach we follow here for deriving the differential-functional equations satisfied by the Poisson generating functions of the first two moments is slightly different from the one we used since the corresponding nonlinear equation for the bivariate generating function $F(z,y) := \sum_{n\ge 0} M_n(y)z^n/n!$ is very involved as given below.

$$\begin{aligned}\frac{\partial}{\partial z}F(z,y)-1 &= F\left(\frac{e^y z}{2},y\right)F\left(\frac{e^{-y}z}{2},y\right)\\ &\quad+\frac{1}{2\pi i}\oint_{|w|=r>0}F\left(\frac{wz}{2},y\right)\left(\frac{F(e^y z/2,y)-w^{-1}e^{-y}F(z/(2w),y)}{w-e^{-y}}\right.\\ &\qquad\left.-\frac{F(e^{-y}z/2,y)-w^{-1}e^{y}F(z/(2w),y)}{w-e^{y}}\right)\mathrm{d}w,\end{aligned}$$

with $F(0,y)=1$.

We use instead a more elementary argument. From the recurrence (61), we obtain, with $\mu_n := \mathbb{E}(X_n)$,

$$\mu_{n+1} = 2^{1-n}\sum_{0\le k\le n}\binom{n}{k}\mu_k + 2^{-n}\sum_{0\le k\le n}\binom{n}{k}|n-2k| \qquad (n\ge 1),$$

the initial condition being $\mu_0 = 0$. Then the Poisson generating function of X_n satisfies the equation

$$\tilde{f}_1(z)+\tilde{f}_1'(z) = 2\tilde{f}_1(z/2)+\tilde{h}_1(z),$$

with $\tilde{f}_1(0)=0$, where $\tilde{h}_1$ is given by

$$\begin{aligned}\tilde{h}_1(z) &= e^{-z}\sum_{n\ge 0}\frac{(z/2)^n}{n!}\sum_{0\le k\le n}\binom{n}{k}|n-2k|\\ &= ze^{-z}\left(I_0(z)+I_1(z)\right),\end{aligned}$$

where we used the identity

$$\sum_{0\le k\le n}\binom{n}{k}|n-2k| = \frac{2n!}{\lfloor n/2\rfloor!(\lceil n/2\rceil-1)!} \qquad (n\ge 1),$$

and $I_\alpha(z)$ denotes the modified Bessel functions

$$I_\alpha(z) := \sum_{n\ge 0}\frac{(z/2)^{2n+\alpha}}{n!\Gamma(n+\alpha+1)}.$$

It is known (see [63]) that, as $|z|\to\infty$,

$$I_\alpha(z) = \begin{cases}\dfrac{e^z}{\sqrt{2\pi z}}\left(1+O(|z|^{-1})\right), & \text{if } |\arg(z)|\le \pi/2-\varepsilon,\\ O\left(|z|^{-1/2}\left(e^{\Re(z)}+e^{-\Re(z)}\right)\right), & \text{if } |\arg(z)|\le\pi,\end{cases}, \tag{65}$$

the O-term holding uniformly in z in each case. Thus, by (65), $\tilde{h}_1 \in \mathscr{JS}$ and

$$\tilde{h}_1(z) = \sqrt{\frac{2z}{\pi}} \left(1 + O(|z|^{-1})\right),$$

for $|z| \to \infty$ in $|\arg(z)| \le \pi/2 - \varepsilon$. Also

$$\mathscr{L}[\tilde{h}_1; s] = (s+2)^{-1/2} s^{-3/2} \qquad (\Re(s) > 0).$$

Thus we can apply the same approach and deduce that

$$\mathbb{E}(X_n) = \frac{n}{\log 2} \sum_{k \in \mathbb{Z}} \frac{G_1(2+\chi_k)}{\Gamma(2+\chi_k)} n^{\chi_k} - \frac{\sqrt{2n}}{\sqrt{2\pi}(\sqrt{2}-1)} + O(1),$$

where $G_1(\omega)$ is the Mellin transform of $\mathscr{L}[\tilde{h}_1; s]/Q(-2s)$

$$G_1(\omega) = \int_0^\infty \frac{s^{\omega - 5/2}}{Q(-2s)\sqrt{s+2}} \,\mathrm{d}s \qquad (\Re(\omega) > 3/2).$$

This proves (63). Numerically, the mean value of the dominant periodic function is $G_1(2)/\log 2 \approx 1.33907\,46494$.

The variance of DPL. Again from (61), we have the recurrence for the second moment $s_n := \mathbb{E}(X_n^2)$

$$s_{n+1} = 2^{-n} \sum_{0 \le k \le n} \binom{n}{k} \left(s_k + s_{n-k} + (n-2k)^2 + 2\mu_k \mu_{n-k} + 4\mu_k |n-2k|\right),$$

for $n \ge 1$ with $s_0 = s_1 = 0$. Since

$$2^{-n} \sum_{0 \le k \le n} \binom{n}{k} (n-2k)^2 = n, \tag{66}$$

we see that the Poisson generating function of s_n satisfies the nonlinear equation

$$\tilde{f}_2(z) + \tilde{f}_2'(z) = 2\tilde{f}_2(z/2) + 2\tilde{f}_1(z/2)^2 + z + 4\tilde{h}_c(z),$$

with $\tilde{f}_2(0) = 0$, where $\tilde{h}_c(z)$ is defined in (62).

Lemma 4.4 *The function $\tilde{h}_c$ is JS-admissible and satisfies*

$$\tilde{h}_c(z) = \tilde{h}_1(z)\tilde{f}_1(z/2) + O(|z|^{1/2}), \tag{67}$$

in the sector $|\arg(z)| \le \pi/2 - \varepsilon$.

Proof: Observe first that

$$\begin{aligned} h_1(z) &= \sum_{k\ge 0} \frac{1}{k!}\left(\frac{z}{2}\right)^k \sum_{n\ge 0} \frac{|n-k|}{n!}\left(\frac{z}{2}\right)^n \\ &= 2\sum_{k\ge 0} \frac{1}{k!}\left(\frac{z}{2}\right)^k \sum_{0\le j\le k} \frac{k-j}{j!}\left(\frac{z}{2}\right)^j \\ &= \frac{2}{2\pi i}\oint_{|w|=r} \frac{e^{z(w+1/w)/2}}{(w-1)^2}\,\mathrm{d}w \qquad (r<1), \end{aligned}$$

since

$$\sum_{0\le j\le k} \frac{k-j}{j!}\left(\frac{z}{2}\right)^j = [w^k]\frac{we^{zw/2}}{(w-1)^2}.$$

On the other hand, since $f_1(z) = \sum_{n\ge 0} \mu_n z^n/n!$, we have, by the same argument,

$$\begin{aligned} h_c(z) &:= e^z \tilde{h}_c(z) \\ &= \sum_{k\ge 0} \frac{\mu_k}{k!}\left(\frac{z}{2}\right)^k \sum_{n\ge 0} \frac{|n-k|}{n!}\left(\frac{z}{2}\right)^n \\ &= \sum_{k\ge 0} \frac{\mu_k}{k!}\left(\frac{z}{2}\right)^k \left(\sum_{0\le n\le k} \frac{k-n}{n!}\left(\frac{z}{2}\right)^n + \sum_{n\ge k} \frac{n-k}{n!}\left(\frac{z}{2}\right)^n\right) \\ &= \sum_{k\ge 0} \frac{\mu_k}{k!}\left(\frac{z}{2}\right)^k \sum_{0\le n\le k} \frac{k-n}{n!}\left(\frac{z}{2}\right)^n + \sum_{n\ge 0} \frac{1}{n!}\left(\frac{z}{2}\right)^n \sum_{0\le k\le n} (n-k)\frac{\mu_k}{k!}\left(\frac{z}{2}\right)^k \\ &= \frac{1}{2\pi i}\oint_{|w|=r<1} f_1\left(\frac{z}{2w}\right)\frac{e^{wz/2}}{(w-1)^2}\,\mathrm{d}w + \frac{1}{2\pi i}\oint_{|w|=r<1} f_1\left(\frac{wz}{2}\right)\frac{e^{z/(2w)}}{(w-1)^2}\,\mathrm{d}w. \end{aligned}$$

To prove condition **(O)**, we start with changes of variables, giving

$$h_c(z) = \frac{z}{2\pi i}\oint_{|w|=|z|}^{\circ} f_1\left(\frac{w}{2}\right)\frac{e^{z^2/(2w)}}{(w-z)^2}\,\mathrm{d}w + \frac{z}{2\pi i}\oint_{|w|=|z|} f_1\left(\frac{w}{2}\right)\frac{e^{z^2/(2w)}}{(w-z)^2}\,\mathrm{d}w,$$

where the first integration circle is indented to the right to avoid the polar singularity $w = z$, and the second to the left. By splitting each integration contour into two parts, we obtain

$$h_c(z) = \frac{z}{2\pi i}\left(\fint + \quad \fint\right) f_1\left(\frac{w}{2}\right)\frac{e^{z^2/(2w)}}{(w-z)^2}\,\mathrm{d}w + O\left(\varepsilon^{-2}\int_{\varepsilon\le|\theta|\le\pi}\left|f_1\left(\frac{|z|e^{i\theta}}{2}\right)\right| e^{|z|(\cos\theta)/2}\,\mathrm{d}\theta\right),$$

where the integration contour $\fint$ is any path connecting the two endpoints $|z|e^{\pm i\varepsilon}$ and indented to the right, and $\fint$ denotes the corresponding symmetric contour with respect to $w = z$ (and indented to the left). Since $\tilde{f}_1 \in \mathscr{J\!S}$, condition **(O)** for $\tilde{h}_c(z)$ is readily checked.

For condition **(I)**, it suffices to prove (67). For that purpose, we use the representation

$$\begin{aligned}\tilde{h}_c(z) &= \frac{e^{-z}}{2\pi i}\oint_{|w|=r<1}\frac{e^{z(w+1/w)/2}}{(w-1)^2}\left(\tilde{f}_1\left(\frac{z}{2w}\right)+\tilde{f}_1\left(\frac{wz}{2}\right)\right)\mathrm{d}w\\
&= \tilde{f}_1\left(\frac{z}{2}\right)\frac{e^{-z}}{2\pi i}\oint_{|w|=r<1}\frac{e^{z(w+1/w)/2}}{(w-1)^2}\,\mathrm{d}w+\frac{e^{-z}}{2\pi i}\oint_{|w|=r<1}e^{z(w+1/w)/2}R_z(w)\,\mathrm{d}w\\
&= \tilde{f}_1(z/2)\tilde{h}_1(z)+\frac{e^{-z}}{2\pi i}\oint_{|w|=1}e^{z(w+1/w)/2}R_z(w)\,\mathrm{d}w,\end{aligned}$$

where

$$\begin{aligned}R_z(w) := \frac{1}{(w-z)^2}\Bigg\{&\left(\tilde{f}_1\left(\frac{z}{2w}\right)-\tilde{f}_1\left(\frac{z}{2}\right)-\tilde{f}_1'\left(\frac{z}{2}\right)\frac{z(w-1)}{2}\right)\\
&+\left(\tilde{f}_1\left(\frac{wz}{2}\right)-\tilde{f}_1\left(\frac{z}{2}\right)+\tilde{f}_1'\left(\frac{z}{2}\right)\frac{z(w-1)}{2}\right)\Bigg\}\end{aligned}$$

is analytic at $w = z$. The error term of $\tilde{h}_c(z) - \tilde{h}_1(z)\tilde{f}_1(z/2)$ can be estimated by a similar argument as that used for checking condition **(O)**. This completes the proof of the Lemma. $\square$

The remaining analysis is now routine. Let $\tilde{V}(z) := \tilde{f}_2(z) - \tilde{f}_1(z)^2 - z\tilde{f}_1'(z)^2$. Then

$$\tilde{V}(z)+\tilde{V}'(z) = 2\tilde{V}(z/2)+\tilde{g}_2(z),$$

where, by Lemma 2.7,

$$\begin{aligned}\tilde{g}_2(z) &= z-\tilde{h}_1(z)^2+4\left(\tilde{h}_c(z)-\tilde{h}_1(z)\tilde{f}_1(z/2)\right)-z\tilde{h}_1'(z)^2-2z\tilde{h}_1'(z)\tilde{f}_1'(z/2)+z\tilde{f}_1''(z)^2\\
&= \left(1-\frac{2}{\pi}\right)z+O(|z|^{1/2}),\end{aligned}$$

for $|\arg(z)| \le \pi/2-\varepsilon$. From this and the analytic properties of the functions involved, we deduce (64).

Remark. The same approach can be extended to more general differential path-length of the form $\sum_{\text{all nodes}} |\mathcal{T}_{\text{left}} - \mathcal{T}_{\text{right}}|^m$ with $m \ge 2$. Interestingly, when $m = 2$, the *mean is identical to the total internal path-length* in view of (66) and the variance is asymptotic to $4n^2$. For $m > 2$, the mean and the variance are asymptotic to

$$\frac{2^{m/2}\Gamma((m+1)/2)}{\sqrt{\pi}(1-2^{1-m})}\,n^{m/2},\qquad \frac{2^m(\Gamma(m+1/2)-\pi^{-1/2}\Gamma((m+1)/2)^2)}{\sqrt{\pi}(1-2^{1-m})}n^m,$$

respectively.

4.4 A weighted path-length (WPL)

Weighted path-lengths of the form $W_n := \sum_{1\le j\le n} w_j \ell_j$ appear often in applications, where ℓ_j denotes the distance of the j-th node (arranged in an appropriate manner, say first level-wise and then left-to-right or in their incoming order) to the root and w_j the weight attached to the j-th node. The calculation of W_n in the case of random DSTs can be carried out recursively by

$$W_{n+1} \stackrel{d}{=} W_{B_n} + W^*_{n-B_n} + \sum_{2\le j\le n+1} w_j,$$

assuming that the root is labelled 1. We consider in this section the case when $w_j = (\log j)^m$, $m \ge 1$. From a technical point of view, it suffices to consider the random variables

$$X_{n+1} \stackrel{d}{=} X_{B_n} + X^*_{n-B_n} + (n+1)(\log(n+1))^m \qquad (n \ge 0),$$

with $X_0 = 0$, since the partial sum $\sum_{2\le j\le n} (\log j)^m$ is nothing but

$$\sum_{2\le j\le n} (\log j)^m = [z^n]\frac{L_{0,m}(z)}{1-z},$$

where

$$L_{a,m}(z) := \sum_{k\ge1} n^{-a}(\log k)^m z^m \qquad (a \ne 1, 2, \dots),$$

on whose analytic properties our analytic approach heavily relies.

The random variables X_n represent the sole example on DSTs we discuss in this paper with non-integral values; they also exhibit an interesting phenomenon in that the mean is of order $n(\log n)^{m+1}$ but the variance is asymptotic to n times a periodic function, in contrast to the orders of DPL.

Theorem 4.5 *The mean and the variance of the weighted path-length X_n are asymptotic to*

$$\mathbb{E}(X_n) = \frac{n(\log n)^{m+1}}{(m+1)\log 2} + n \sum_{1\le j\le m} c_{m,j}(\log n)^j + nP_{w,\mu}(\log_2 n) + O\left((\log n)^{m+1}\right),$$
$$\mathbb{V}(X_n) = nP_{w,\sigma}(\log_2 n) + O\left((\log n)^{2m+2}\right),$$

respectively, where the $c_{m,j}$'s are constants depending on m, and $P_{w,\mu}$ and $P_{w,\sigma}$ are 1-periodic, smooth functions.

That the variance is linear is well-predicted by the deep theorem of Schachinger derived in [58] since the second difference of the sequence $n(\log n)^m$ is $o(n^{-1/2-\varepsilon})$. Our approach has the advantage of providing more precise approximations.

The new ingredient we need is incorporated in the following lemma.

Lemma 4.6 ([21]) *The function $L_{a,m}(z)$ can analytically be continued into the cut-plane $\mathbb{C} \setminus [1, \infty)$ with a sole singularity at $z = 1$ near which it admits the asymptotic approximation*

$$L_{a,m}(e^{-s}) = \Gamma(1-a)s^{a-1}(-\log s)^m + O(1),$$

the O-term holding uniformly for $|\arg(s)| \le \pi - \varepsilon$.

Indeed, the tools developed in [21] can also be easily extended to similar "toll-functions" such as nH_n^m. Details are left for the interested readers.

5 Conclusions and extensions

We showed in this paper, through many shape parameters on random DSTs that the crucial use of the normalization $\tilde{V}(z) := \tilde{f}_2(z) - \tilde{f}_1(z)^2 - z\tilde{f}'(z)^2$ at the level of Poisson generating function is extremely helpful in simplifying the asymptotic analysis of the variance as well as the resulting expressions. The same idea can be applied to a large number of concrete problems with a binomial splitting procedure. These and some related topics and extensions will be pursued elsewhere. We briefly mention in this final section some extensions and related properties.

Central limit theorems. All shape parameters we considered in this paper are asymptotically normally distributed in the sense of convergence in distribution. We describe the results in this section and merely indicate the methods of proofs. The only case that requires a separate study is NPL of random b-DSTs with $b \geq 2$ (a bivariate consideration of the limit laws is needed), details being given in a future paper.

Theorem 5.1 *The internal path-length, the peripheral path-length, the number of leaves, the differential path-length, the weighted path-length of random DSTs, and the key-wise path-length of random b-DSTs with $b \geq 2$ are all asymptotically normally distributed*

$$\frac{X_n - \mathbb{E}(X_n)}{\sqrt{\mathbb{V}(X_n)}} \xrightarrow{d} \mathscr{N}(0,1),$$

where X_n denotes any of these shape parameters, $\xrightarrow{d}$ stands for convergence in distribution, and $\mathscr{N}(0,1)$ is a standard normal distribution with zero mean and unit variance.

See Figure 10 for a plot of the histograms of DPL.

The method of moments applies to all these cases and establishes the central limit theorems; similar details are given as in [31] (the asymptotic normality of the number of leaves being already proved there as a special case).

In a parallel way, contraction method also works well for all these shape parameters; see [51, 52, 53].

On the other hand, Schachinger's asymptotic normality results cover the IPL, PPL, number of leaves and WPL, but not PPL and KPL on b-DSTs, although his approach may be modified for that purpose.

Finally, the complex-analytic approach used in [35] for internal path-length may be extended to prove some of these cases, but the proofs are messy, although the results established are often stronger (for example, with convergence rate).

The depth. The asymptotic analysis we used in this paper can also be extended to the depth (the distance between a randomly chosen internal node and the root) although it is of logarithmic order. Let X_n denote the depth of a random DST of n nodes. The starting point is to consider the expected profile polynomial

$$P_n(y) := \sum_{0 \leq k < n} n\mathbb{P}(X_n = k) y^k,$$

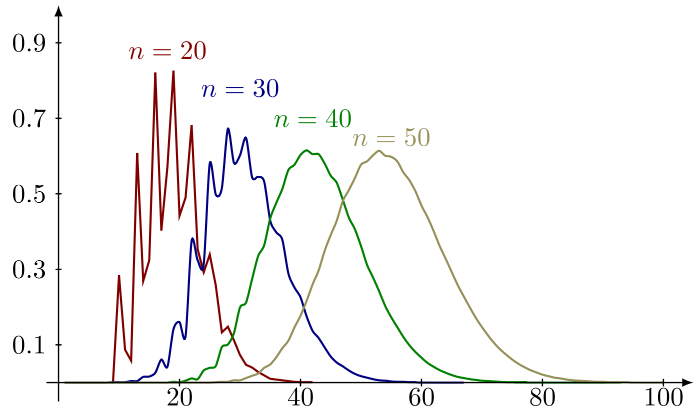


Fig. 10: *The histograms of DPL for* $n = 20, 30, 40$ *and* 50*, normalized by their standard deviations.*

where $n\mathbb{P}(X_n = k)$ is nothing but the expected number of internal nodes at distance k to the root. Then we have the recurrence

$$P_{n+1}(y) = 1 + y2^{-n} \sum_{0 \le k \le n} \binom{n}{k} (P_k(y) + P_{n-k}(y)) \qquad (n \ge 0),$$

with $P_0(y) = 0$. From this relation, we obtain the equation for the Poisson generating function $\tilde{F}(z, y)$ of $P_n(y)$

$$\tilde{F}(z, y) + \frac{\partial}{\partial z}\tilde{F}(z, y) = 2y\tilde{F}\left(\frac{z}{2}, y\right) + 1,$$

with $\tilde{F}(0, y) = 0$. It follows, by taking coefficients of z^n on both sides and by solving the resulting recurrence, that

$$P_n(y) = \sum_{1 \le k \le n} \binom{n}{k} (-1)^{k-1} \prod_{0 \le j \le k-2} \left(1 - \frac{y}{2^j}\right) \qquad (n \ge 1);$$

see [44, p. 504] for a different proof. Asymptotic approximation to $P_n(y)$ can then be obtained by Rice's integral formula

$$P_n(y) = n - \frac{y-1}{2\pi i} \int_{(3/2)} \frac{\Gamma(n+1)\Gamma(-s)Q(y)}{\Gamma(n+1-s)(1-2^{1-s}y)Q(2^{1-s}y)} \,\mathrm{d}s,$$

for $|y - 1| \le \varepsilon$. More precisely, if $t \in \mathbb{C}$ lies in a small neighborhood of the origin, then

$$\begin{aligned} \mathbb{E}(e^{X_n t}) &= \frac{P_n(e^t)}{n} \\ &= \frac{(e^t - 1)Q(e^t)}{Q(1)\log 2} \sum_{k \in \mathbb{Z}} \Gamma\left(-1 - \frac{t}{\log 2} - \chi_k\right) n^{t/\log_2 + \chi_k} \left(1 + O\left(n^{-1}\right)\right) + O(n^{-1}), \end{aligned} \qquad (68)$$

uniformly for $|t| \le \varepsilon$. Alternatively, one can also apply the Laplace-Mellin-de-Poissonization approach and obtain the same type of result for not only DSTs but also for more general b-DSTs. See [48, 49] for a more general and detailed treatment (by a different approach).

The estimate (68) leads to effective asymptotic estimates for all moments of $X_n - \log_2 n$ by standard arguments; see [32]. In particular, we obtain

$$\begin{aligned}\mathbb{E}(X_n) &= \log_2 n + \frac{\gamma-1}{\log 2} + \frac{1}{2} - \sum_{k\ge1}\frac{1}{2^k-1} + \varpi_1(\log_2 n) + O\left(n^{-1}\log n\right),\\ \mathbb{V}(X_n) &= \frac{1}{12} + \frac{1}{(\log 2)^2}\left(1+\frac{\pi^2}{6}\right) - \sum_{k\ge1}\frac{2^k}{(2^k-1)^2} + \varpi_5(\log_2 n) + O\left(n^{-1}\log^2 n\right),\end{aligned}$$

where the estimate for the mean is exactly (7) with ϖ_1 given in (8) and ϖ_5 is a smooth periodic function.

An analytic extension. From a purely analytic viewpoint, the underlying differential-functional equation (13) for the moments can be extended to an equation of the form

$$\sum_{0\le j\le b}\binom{b}{j}\tilde{f}^{(j)}(z) = \alpha\tilde{f}\left(\frac{z}{\beta}\right) + \tilde{g}(z) \qquad (\alpha>0;\beta>1),$$

for which our approach still applies, leading to the functional equation for the Laplace transform

$$(s+1)^b\mathscr{L}[\tilde{f};s] = \alpha\beta\mathscr{L}[\tilde{f};\beta s] + \mathscr{L}[\tilde{g};s].$$

The natural normalizing function is then provided by

$$Q_\beta(-s) := \prod_{j\ge1}\left(1+\frac{s}{\beta^j}\right)^b,$$

and the corresponding Laplace-Mellin asymptotic analysis is similar.

In particular, the case when $\alpha=\beta=m$ corresponds to a straightforward extension of binary DSTs to m-ary DSTS (and the binary unbiased Bernoulli random variable to the uniform distribution over $\{0,1,\dots,m-1\}$). The stochastic behaviors of all shape parameters on such trees follow the same patterns as showed in this paper.

Yet another concrete instance arises in the so-called Eden model studied by Dean and Majumdar [10], which corresponds to $\alpha=m$ and $\beta>1$. The model is constructed in the following way. We start at time $t=0$ at which we have an empty node. Then at time $t=T$, where $T\sim$ Exponential(1), we fill the empty node and attach to it m different empty nodes. The process then continues independently for each empty node by the following recursive rule. Once an empty node of depth j is attached to a tree at time $t=t'$, it is then filled at time point $t'+T$, where $T\sim E(\beta^j)$, and m new empty nodes are attached to it.

The mean and the variance of the number of filled nodes at a large time of such trees are studied in details in [10]. Since the model is continuous, there is no need to de-Poissonize to derive the asymptotics of the coefficient; as a consequence, no correction term as we used in this paper is required for the asymptotics of the variance.

Other DST-type recurrences. While the technique of Poissonized variance with correction remains useful for the natural case when the Bernoulli random variable is no longer symmetric, the Laplace-Mellin approach does not apply directly. Other asymptotic ingredients are needed such as a direct manipulation of the Mellin transforms; see [49] and the references therein.

DST-type structures and recurrences also arise in other statistical physical models such as the diffusion-limited aggregation; see [1, 5].

Acknowledgement

We thank the referee for opportune helpful comments and the more precise title.

Appendix. An Elementary Approach to the Asymptotic Linearity of the Variance.

We describe briefly here a direct elementary approach to the variance of random variables satisfying the recurrence

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

where

$$\pi_{n,k} := \mathbb{P}(B_n = k) = \binom{n}{k} 2^{-n} \qquad (0 \le k \le n).$$

The starting point is to consider the recurrence satisfied by the variance $v_n := \mathbb{V}(X_n)$

$$v_{n+1} = \sum_{0 \le k \le n} \pi_{n,k} \left(v_k + v_{n-k} \right) + u_n + \mathbb{V}(T_n),$$

where $\mu_k := \mathbb{E}(X_n)$ and

$$u_n := \sum_{0 \le k \le n} \pi_{n,k} \left(\mu_k + \mu_{n-k} - \mu_{n+1} + \mathbb{E}(T_n) \right)^2.$$

In most cases, we have the estimate $\mu_k = \tilde{f}_1(k) + O(k^{\varepsilon})$. This, together with the Gaussian approximation of the binomial distribution, implies that

$$\begin{aligned} u_n &\approx \sum_{\substack{|k-n/2| = o(n^{2/3}) \\ k = n/2 + x\sqrt{n}/2}} \pi_{n,k} \left(\tilde{f}_1\left(\frac{n}{2} + \frac{x}{2}\sqrt{n} \right) + \tilde{f}_1\left(\frac{n}{2} - \frac{x}{2}\sqrt{n} \right) - \tilde{f}_1(n+1) + \mathbb{E}(T_n) \right)^2 \\ &\approx \sum_{\substack{|k-n/2| = o(n^{2/3}) \\ k = n/2 + x\sqrt{n}/2}} \pi_{n,k} \left(2\tilde{f}_1\left(\frac{n}{2} \right) - \tilde{f}_1(n) - \tilde{f}_1'(n) + \mathbb{E}(T_n) \right)^2 \\ &\approx \left(2\tilde{f}_1\left(\frac{n}{2} \right) - \tilde{f}_1(n) - \tilde{f}_1'(n) + \mathbb{E}(T_n) \right)^2. \end{aligned}$$

But then (see (13) below)

$$2\tilde{f}_1\left(\frac{n}{2} \right) - \tilde{f}_1(n) - \tilde{f}_1'(n) + \mathbb{E}(T_n) = \mathbb{E}(T_n) - \tilde{h}_1(n),$$

where

$$\tilde{h}_1(z) := e^{-z} \sum_{j \ge 0} \frac{\mathbb{E}(T_j)}{j!} z^j.$$

The order of the difference $\mathbb{E}(T_n) - \tilde{h}_1(n) \approx n|\tilde{h}_1''(n)|$ are expected to be small, roughly $O(n^{\varepsilon})$ in all cases we consider here. Consequently, the variance is asymptotically linear; see [31, 58] for more precise details.

We see clearly that the smallness of the variance results naturally from the high concentration of the binomial distribution near its mean.

References

[1] D. Aldous and P. Shields. A diffusion limit for a class of randomly-growing binary trees. *Probab. Theory Related Fields*, 79(4):509–542, 1988.

[2] Z.-D. Bai, H.-K. Hwang, W.-Q. Liang, and T.-H. Tsai. Limit theorems for the number of maxima in random samples from planar regions. *Electron. J. Probab.*, 6:no. 3, 41 pp. (electronic), 2001.

[3] B. C. Berndt. *Ramanujan's notebooks. Part I.* Springer-Verlag, New York, 1985. With a foreword by S. Chandrasekhar.

[4] M. G. B. Blum, O. François, and S. Janson. The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *Ann. Appl. Probab.*, 16(4):2195–2214, 2006.

[5] R. M. Bradley and P. N. Strenski. Directed aggregation on the bethe lattice: Scaling, mappings, and universality. *Phys. Rev. B*, 31(7):4319–4328, Apr 1985.

[6] W.-M. Chen and H.-K. Hwang. Analysis in distribution of two randomized algorithms for finding the maximum in a broadcast communication model. *J. Algorithms*, 46(2):140–177, 2003.

[7] H.-H. Chern, M. Fuchs, and H.-K. Hwang. Phase changes in random point quadtrees. *ACM Trans. Algorithms*, 3(2):Art. 12, 51, 2007.

[8] H.-H. Chern, H.-K. Hwang, and T.-H. Tsai. An asymptotic theory for Cauchy-Euler differential equations with applications to the analysis of algorithms. *J. Algorithms*, 44(1):177–225, 2002. Analysis of algorithms.

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

[10] D. S. Dean and S. N. Majumdar. Phase transition in a generalized Eden growth model on a tree. *J. Stat. Phys.*, 124(6):1351–1376, 2006.

[11] F. Dennert and R. Grübel. Renewals for exponentially increasing lifetimes, with an application to digital search trees. *Ann. Appl. Probab.*, 17(2):676–687, 2007.

[12] L. Devroye. A study of trie-like structures under the density model. *Ann. Appl. Probab.*, 2(2):402–434, 1992.

[13] L. Devroye. Universal limit laws for depths in random trees. *SIAM J. Comput.*, 28(2):409–432 (electronic), 1999.

[14] M. Drmota. The variance of the height of digital search trees. *Acta Inform.*, 38(4):261–276, 2002.

[15] M. Drmota. *Random trees*. SpringerWienNewYork, Vienna, 2009. An interplay between combinatorics and probability.

[16] M. Drmota, B. Gittenberger, A. Panholzer, H. Prodinger, and M. D. Ward. On the shape of the fringe of various types of random trees. *Math. Methods Appl. Sci.*, 32(10):1207–1245, 2009.

[17] M. Drmota and W. Szpankowski. (Un)expected behavior of digital search tree profile. In *SODA*, pages 130–138, 2009.

[18] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. Tricomi. *Higher transcendental functions. Vol. I.* McGraw-Hill(New York), 1953.

[19] G. Fayolle, P. Flajolet, and M. Hofri. On a functional equation arising in the analysis of a protocol for a multi-access broadcast channel. *Adv. in Appl. Probab.*, 18(2):441–472, 1986.

[20] R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Philosophical Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

[21] P. Flajolet. Singularity analysis and asymptotics of Bernoulli sums. *Theoret. Comput. Sci.*, 215(1-2):371–381, 1999.

[22] P. Flajolet, X. Gourdon, and P. Dumas. Mellin transforms and asymptotics: harmonic sums. *Theoret. Comput. Sci.*, 144(1-2):3–58, 1995. Special volume on mathematical analysis of algorithms.

[23] P. Flajolet and A. Odlyzko. Singularity analysis of generating functions. *SIAM J. Discrete Math.*, 3(2):216–240, 1990.

[24] P. Flajolet and B. Richmond. Generalized digital trees and their difference-differential equations. *Random Structures Algorithms*, 3(3):305–320, 1992.

[25] P. Flajolet and N. Saheb. The complexity of generating an exponentially distributed variate. *J. Algorithms*, 7(4):463–488, 1986.

[26] P. Flajolet and R. Sedgewick. Digital search trees revisited. *SIAM J. Comput.*, 15(3):748–767, 1986.

[27] P. Flajolet and R. Sedgewick. Mellin transforms and asymptotics: finite differences and Rice's integrals. *Theoret. Comput. Sci.*, 144(1-2):101–124, 1995. Special volume on mathematical analysis of algorithms.

[28] P. Flajolet and R. Sedgewick. *Analytic combinatorics*. Cambridge University Press, Cambridge, 2009.

[29] A. Hald. *On the history of series expansions of frequency functions and sampling distributions, 1873–1944*. Matematisk-Fysiske Meddelelser. 49. Copenhagen: The Royal Danish Academy of Sciences and Letters., 2002.

[30] F. Hubalek. On the variance of the internal path length of generalized digital trees – the Mellin convolution approach. *Theoret. Comput. Sci.*, 242(1-2):143–168, 2000.

[31] F. Hubalek, H.-K. Hwang, W. Lew, H. Mahmoud, and H. Prodinger. A multivariate view of random bucket digital search trees. *J. Algorithms*, 44(1):121–158, 2002. Analysis of algorithms.

[32] H.-K. Hwang. On convergence rates in the central limit theorems for combinatorial structures. *European J. Combin.*, 19(3):329–343, 1998.

[33] P. Jacquet and E. Merle. Analysis of a stack algorithm for csma-cd random length packet communication. *IEEE Transactions on Information Theory*, 36(2):420–426, 1990.

[34] P. Jacquet and M. Régnier. Normal limiting distribution of the size of tries. In *Performance'87 (Brussels, 1987)*, pages 209–223. North-Holland, Amsterdam, 1988.

[35] P. Jacquet and W. Szpankowski. Asymptotic behavior of the Lempel-Ziv parsing scheme and [in] digital search trees. *Theoret. Comput. Sci.*, 144(1-2):161–197, 1995. Special volume on mathematical analysis of algorithms.

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

[37] P. Jacquet, W. Szpankowski, and J. Tang. Average profile of the Lempel-Ziv parsing scheme for a Markovian source. *Algorithmica*, 31(3):318–360, 2001. Mathematical analysis of algorithms.

[38] S. Janson. Rounding of continuous random variables and oscillatory asymptotics. *Ann. Probab.*, 34(5):1807–1826, 2006.

[39] P. Kirschenhofer and H. Prodinger. Eine Anwendung der Theorie der Modulfunktionen in der Informatik. *Österreich. Akad. Wiss. Math.-Natur. Kl. Sitzungsber. II*, 197(4-7):339–366, 1988.

[40] P. Kirschenhofer and H. Prodinger. Further results on digital search trees. *Theoret. Comput. Sci.*, 58(1-3):143–154, 1988. Thirteenth International Colloquium on Automata, Languages and Programming (Rennes, 1986).

[41] P. Kirschenhofer and H. Prodinger. On some applications of formulae of Ramanujan in the analysis of algorithms. *Mathematika*, 38(1):14–33, 1991.

[42] P. Kirschenhofer, H. Prodinger, and W. Szpankowski. Digital search trees again revisited: the internal path length perspective. *SIAM J. Comput.*, 23(3):598–616, 1994.

[43] 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):923–964 (electronic), 2000.

[44] D. E. Knuth. *The art of computer programming. Volume 3: Sorting and searching*. Addison-Wesley Publishing Co., Reading, Mass., second edition, 1998.

[45] A. G. Konheim and D. J. Newman. A note on growing binary trees. *Discrete Math.*, 4:57–63, 1973.

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

[47] G. Louchard. Digital search trees revisited. *Cahiers Centre Études Rech. Opér.*, 36:259–278, 1994. Hommage à Simone Huyberechts.

[48] G. Louchard and W. Szpankowski. Average profile and limiting distribution for a phrase size in the Lempel-Ziv parsing algorithm. *IEEE Trans. Inform. Theory*, 41(2):478–488, 1995.

[49] G. Louchard, W. Szpankowski, and J. Tang. Average profile of the generalized digital search tree and the generalized Lempel-Ziv algorithm. *SIAM J. Comput.*, 28(3):904–934 (electronic), 1999.

[50] H. M. Mahmoud. *Evolution of random search trees*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons Inc., New York, 1992. A Wiley-Interscience Publication.

[51] R. Neininger. On a multivariate contraction method for random recursive structures with applications to Quicksort. *Random Structures Algorithms*, 19(3-4):498–524, 2001. Analysis of algorithms (Krynica Morska, 2000).

[52] R. Neininger and L. Rüschendorf. A general limit theorem for recursive algorithms and combinatorial structures. *Ann. Appl. Probab.*, 14(1):378–418, 2004.

[53] R. Neininger and L. Rüschendorf. A survey of multivariate aspects of the contraction method. *Discrete Math. Theor. Comput. Sci.*, 8(1):31–56 (electronic), 2006.

[54] F. W. J. Olver. *Asymptotics and special functions*. Academic Press, 1974. Computer Science and Applied Mathematics.

[55] B. Pittel. Paths in a random digital tree: limiting distributions. *Adv. in Appl. Probab.*, 18(1):139–155, 1986.

[56] H. Prodinger. External internal nodes in digital search trees via Mellin transforms. *SIAM J. Comput.*, 21(6):1180–1183, 1992.

[57] H. Prodinger. Hypothetical analyses: approximate counting in the style of Knuth, path length in the style of Flajolet. *Theoret. Comput. Sci.*, 100(1):243–251, 1992.

[58] W. Schachinger. On the variance of a class of inductive valuations of data structures for digital search. *Theoret. Comput. Sci.*, 144(1-2):251–275, 1995. Special volume on mathematical analysis of algorithms.

[59] W. Schachinger. Asymptotic normality of recursive algorithms via martingale difference arrays. *Discrete Math. Theor. Comput. Sci.*, 4(2):363–397 (electronic), 2001.

[60] W. Szpankowski. The evaluation of an alternative sum with applications to the analysis of some data structures. *Inform. Process. Lett.*, 28(1):13–19, 1988.

[61] W. Szpankowski. A characterization of digital search trees from the successful search viewpoint. *Theoret. Comput. Sci.*, 85(1, Algorithms Automat. Complexity Games):117–134, 1991.

[62] W. Szpankowski. *Average case analysis of algorithms on sequences*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2001. With a foreword by Philippe Flajolet.

[63] E. T. Whittaker and G. N. Watson. *A course of modern analysis. An introduction to the general theory of infinite processes and of analytic functions; with an account of the principal transcendental functions*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, fourth edition, 1927.