Dependence between path-length and size in random digital trees

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

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

January 24, 2017

Abstract

We study the size and the external path length of random tries and show that they are asymptotically independent in the asymmetric case but strongly dependent with small periodic fluctuations in the symmetric case. Such an unexpected behavior is in sharp contrast to the previously known results on random tries that the size is totally positively correlated to the internal path length and that both tend to the same normal limit law. These two dependence examples provide concrete instances of bivariate normal distributions (as limit laws) whose correlation is 0, 1 and periodically oscillating. Moreover, the same type of behaviors is also clarified for other classes of digital trees such as bucket digital trees and Patricia tries.

AMS 2010 Subject Classifications. 60C05 60F05 68P05 05C05 68W40
Keywords. Random tries, covariance, total path length, Pearson's correlation coefficient, asymptotic normality, poissonization/de-Poissonization, integral transform, contraction method.

1 Introduction

Tries are one of the most fundamental tree-type data structures in computer algorithms; see Knuth [18] and Mahmoud [19] for a general introduction. Their general efficiency depends on several shape parameters, the principal ones including the depth, the height, the size, the internal path-length (IPL), and the external path-length (EPL); see below for a more precise description of those studied in this paper. While most of these measures have been extensively investigated in the literature, we are concerned here with the question: *how does the EPL depend on the size in a random trie?* Surprisingly, while the pair $(\mathrm{IPL}, \mathrm{size})$ is known to have asymptotic correlation coefficient tending to one and to have the same normal limit law after each being properly normalized (see [10, 12]), this paper aims to show that the pair $(\mathrm{EPL}, \mathrm{size})$ exhibits a completely different behavior depending on the parameter of the underlying random bits being biased or unbiased. This is a companion paper to [2] where we clarified the dependence structure of another class of search trees in computer algorithms.

Given a sequence of binary strings (or keys), one can construct a binary trie (very similar to constructing a dictionary of binary words) as follows. If $n = 1$, then the trie consists of a single root-node holding the sole string; if $n \geqslant 2$, the root is used to direct the strings into the corresponding subtree: if the first bit of the input string is 0 (or 1), then the string goes to the left (or right) subtree; strings directed to the same subtree are then processed recursively in the same manner but instead of splitting according to the first bit, the second bit of each string is then used. In this way, a binary dictionary-type tree with two types of nodes is constructed: external nodes for storing strings and internal nodes for splitting the strings; see Figure 1 for a trie of seven strings.

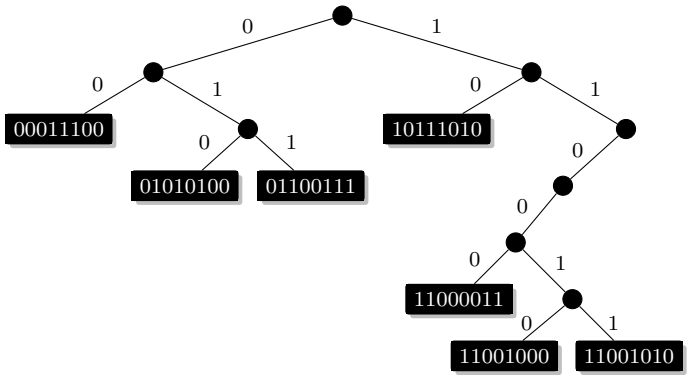

Figure 1: *A trie with $n = 7$ records: the (filled) circles represent internal nodes and rectangles holding the binary strings are external nodes. In this example, $S_n = 8$, $K_n = 27$, and $N_n = 18$.*

The random trie model we consider here assumes that each of the n binary keys is an infinite sequence of independent Bernoulli bits each with success probability $0 < p < 1$. Then the trie constructed from this sequence is a random trie.

We define three shape parameters in a random trie of n strings:

- Size S_n: the total number of internal nodes used (the circle nodes in Figure 1);
- IPL (or node path-length, NPL) N_n: the sum of the distance between the root and each internal node;
- EPL (or key path-length, KPL) K_n: the sum of the distance between the root and each external node.

We will use mostly NPL in place of IPL, and KPL in place of EPL, the reason being an easier comparison with the corresponding results derived for random m-ary search trees in the companion paper [2]; see below for more details.

By the recursive definition and our model assumption, we have the following recurrence relations

$$\begin{cases} S_n \stackrel{d}{=} S_{B_n} + S^*_{n-B_n} + 1, \\ K_n \stackrel{d}{=} K_{B_n} + K^*_{n-B_n} + n, & (n \geqslant 2), \\ N_n \stackrel{d}{=} N_{B_n} + N^*_{n-B_n} + S_{B_n} + S^*_{n-B_n}, \end{cases} \tag{1}$$

with the initial conditions $S_n = K_n = N_n = 0$ for $n \leqslant 1$, where $B_n = \mathrm{Binom}(n, p)$ denotes a binomial distribution with parameters n and $p \in (0, 1)$. Also $(S_n^*), (K_n^*)$, and (N_n^*) are independent copies of $(S_n), (K_n)$ and (N_n), respectively. While many stochastic properties of these random variables are known (see Clément et al. [3], Devroye [5] and [10] and many references cited there), much less attention has been paid to their correlation and dependence structure.

The asymptotic behaviors of the moments of random variables defined on tries typically depend on the ratio $\frac{\log p}{\log q}$ being rational or irrational, where $q = 1 - p$. So we introduce, similar to [10], the notation

$$\mathscr{F}[g](z) = \begin{cases} \sum_{k\in\mathbb{Z}} g_k z^{-\chi_k}, & \text{if } \frac{\log p}{\log q} \in \mathbb{Q}; \\ g_0, & \text{if } \frac{\log p}{\log q} \notin \mathbb{Q}, \end{cases} \tag{2}$$

where g_k represents a sequence of (Fourier) coefficients and $\chi_k = \frac{2rk\pi i}{\log p}$ when $\frac{\log p}{\log q} = \frac{r}{l}$ with r and l coprime. In simpler words, $\mathscr{F}[g](z)$ is a periodic function in the rational case, and a constant in the irrational case. We also use $\mathscr{F}[\cdot](z)$ as a generic symbol if the exact form of the underlying sequence matters less, and in this case each occurrence may not represent the same function.

With this notation, the asymptotics of the mean and the variance are summarized in the following table; see [10, 15, 19] and the references therein for more information.

Shape parameters	$\frac{1}{n}$(mean) $\sim$	$\frac{1}{n}$(variance) $\sim$
Size S_n	$\mathscr{F}[\cdot](n)$	$\mathscr{F}[g^{(1)}](n)$
NPL N_n	$\frac{\mathbb{E}(S_n)}{n} \cdot \frac{\log n}{h}$	$\frac{\mathbb{V}(S_n)}{n} \cdot \frac{(\log n)^2}{h^2}$
KPL K_n	$\frac{\log n}{h} + \mathscr{F}[\cdot](n)$	$\frac{pq \log^2 \frac{p}{q}}{h^2} \cdot \frac{\log n}{h} + \mathscr{F}[g^{(3)}](n)$
Depth D_n	$\mathbb{E}(D_n) = \frac{\mathbb{E}(K_n)}{n}$	$\mathbb{V}(D_n) = \frac{\mathbb{V}(K_n)}{n} + O(1)$

Table 1: *Asymptotic patterns of the means and the variances of the shape parameters discussed in this paper. Here $\mathscr{F}[\cdot](n)$ differs from one occurrence to another and $h = -p\log p - q\log q$ denotes the entropy. Expressions for $g_k^{(1)}$ and $g_k^{(3)}$ will be given below. All three random variables S_n, N_n, K_n are asymptotically normally distributed.*

Note specially that the leading constant

$$\lambda = \lambda_p := \frac{pq \log^2 \frac{p}{q}}{h^3} = \frac{(p\log^2 p + q \log^2 q) - h^2}{h^3}$$

in the asymptotic approximation to $\mathbb{V}(K_n)$ equals zero when $p = q$, implying that $\mathbb{V}(K_n)$ is not of order $n \log n$ but of linear order in the symmetric case. *This change of order can be regarded as the source property distinguishing between the dependence and independence of K_n on S_n.*

On the other hand, we have the relation $\mathbb{E}(K_n) = \mathbb{E}(D_n)n$ between the external path length and the depth D_n, which is defined to be the distance between the root and a randomly chosen external node (each with the same probability). Furthermore, we also have the asymptotic equivalent $\mathbb{V}(K_n) \sim \mathbb{V}(D_n)n$ when $p \neq 1/2$ (or $\lambda > 0$), and a central limit theorem for D_n; see Devroye [4].

From Table 1, we see roughly that each internal node contributes $\frac{\log n}{h}$ to N_n, namely, that $N_n \approx S_n \cdot \frac{\log n}{h}$. Indeed, it was proved in [10] that the correlation coefficient of S_n and N_n satisfies

$$\rho(S_n, N_n) \sim 1 \qquad (0 < p < 1). \tag{3}$$

Such a linear correlation was further strengthened in [12], where it was proved that both random variables tend to the *same* normal limit law $\mathcal{N}_1$ (with zero mean and unit variance)

$$\left(\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathbb{V}(S_n)}}, \frac{N_n - \mathbb{E}(N_n)}{\sqrt{\mathbb{V}(N_n)}} \right) \xrightarrow{d} (\mathcal{N}_1, \mathcal{N}_1),$$

where $\xrightarrow{d}$ denotes convergence in distribution. In terms of the bivariate normal law $\mathcal{N}_2$ (see Tong [27]), we can write

$$\left(\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathbb{V}(S_n)}}, \frac{N_n - \mathbb{E}(N_n)}{\sqrt{\mathbb{V}(N_n)}} \right)^{\intercal} \xrightarrow{d} \mathcal{N}_2(0, E_2),$$

where $E_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ is a singular matrix and $\mathbf{A}^{\intercal}$ denotes the transpose of matrix $\mathbf{A}$.

We show that the correlation and dependence of K_n on S_n are drastically different. We start with their correlation coefficient.

Theorem A. *The covariance of the number of internal nodes and KPL in a random trie of n strings satisfies*

$$\mathrm{Cov}(S_n, K_n) \sim n \mathscr{F}[g^{(2)}](n),$$

where $g_k^{(2)}$ is given in Proposition A below, and their correlation coefficient satisfies

$$\rho(S_n, K_n) \sim \begin{cases} 0, & \text{if } p \neq \frac{1}{2} \\ F(n), & \text{if } p = \frac{1}{2}. \end{cases} \tag{4}$$

Here $F(n) = \frac{\mathscr{F}[g^{(2)}](n)}{\sqrt{\mathscr{F}[g^{(1)}](n)\mathscr{F}[g^{(3)}](n)}}$ is a periodic function with average value $0.927\cdots$.

The result (4) is to be compared with (3) (which holds for all $p \in (0, 1)$): *the surprising difference here comes not only from the (common) distinction between $p = \frac{1}{2}$ and $p \neq \frac{1}{2}$ but also from the (less expected) intrinsic asymptotic nature.*

Furthermore, we show that this different behavior cannot be ascribed to the weak measurability of nonlinear dependence of Pearson's correlation coefficient because the limiting distribution also exhibits a similar dependence pattern. (For the univariate central limit theorems implied by the result below, see Jacquet and Régnier [14] where such results were first established.)

Theorem B. *(i) For $p \neq \frac{1}{2}$, we have*

$$\left(\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathbb{V}(S_n)}}, \frac{K_n - \mathbb{E}(K_n)}{\sqrt{\mathbb{V}(K_n)}} \right)^{\intercal} \xrightarrow{d} \mathcal{N}_2(0, I_2),$$

where I_2 denotes the 2×2 identity matrix.

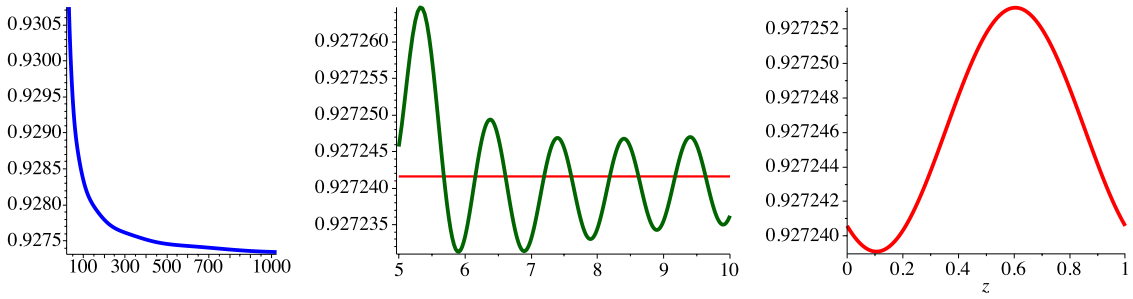

Figure 2: $p = \frac{1}{2}$: *periodic fluctuations of (i)* $\rho(S_n, K_n)$ *(left) for* $n = 32, \ldots, 1024$*, (ii)* $\frac{\mathrm{Cov}(S_n,K_n)}{\sqrt{\mathbb{V}(S_n)(\mathbb{V}(K_n)+1.046)}}$ *(middle) in logarithmic scale, and (iii)* $F(n)$ *by its Fourier series expansion (right). Note that the fluctuations are only visible by a proper correction in the denominator because the amplitude of* F *is very small:* $|F(\cdot)| \leqslant 1.5 \times 10^{-5}$.

(ii) For $p = \frac{1}{2}$*, we have*

$$\Sigma_n^{-\frac{1}{2}} \begin{pmatrix} S_n - \mathbb{E}(S_n) \\ K_n - \mathbb{E}(K_n) \end{pmatrix} \xrightarrow{d} \mathcal{N}_2(0, I_2),$$

where Σ_n *denotes the (asymptotic) covariance matrix of* S_n *and* K_n*:*

$$\Sigma_n := n \begin{pmatrix} \mathscr{F}[g^{(1)}](n) & \mathscr{F}[g^{(2)}](n) \\ \mathscr{F}[g^{(2)}](n) & \mathscr{F}[g^{(3)}](n) \end{pmatrix}.$$

Alternatively, we may define

$$\Sigma_n := n \begin{pmatrix} \mathscr{F}[g^{(1)}](n) & \mathscr{F}[g^{(2)}](n) \\ \mathscr{F}[g^{(2)}](n) & \lambda \log n + \mathscr{F}[g^{(3)}](n) \end{pmatrix}.$$

Then both cases can be stated in one as

$$\Sigma_n^{-\frac{1}{2}} \begin{pmatrix} S_n - \mathbb{E}(S_n) \\ K_n - \mathbb{E}(K_n) \end{pmatrix} \xrightarrow{d} \mathcal{N}_2(0, I_2).$$

On the other hand, since for bivariate normal distribution, zero correlation implies independence (see [27]), it is more transparent to split the statement into two cases. See Figure 3 for (Monte Carlo) 3D-plots of the joint distributions of (S_n, K_n) when $n = 10^7$.

These results are to be compared with the corresponding ones for random m-ary search trees [2], and the differences for correlation coefficients are summarized in Table 2. Furthermore, the joint distribution for m-ary search trees undergoes a phase change at $m = 26$: if the branching factor m satisfies $3 \leqslant m \leqslant 26$, then the space requirement is asymptotically independent with the KPL and NPL, while for $m \geqslant 27$, their limiting joint distributions contain periodic fluctuations and are dependent; see [2] for more information.

The dependence phenomena as those discovered in this paper are not limited to random tries and have indeed a wider range of connections. They also appear in different forms in other structures and algorithms with an underlying binomial splitting process; see Flajolet [6] and [10, 13] for references on data structures, algorithms, conflict resolution protocols and stochastic models. A typical example is the dependence between the number of coin-tossings

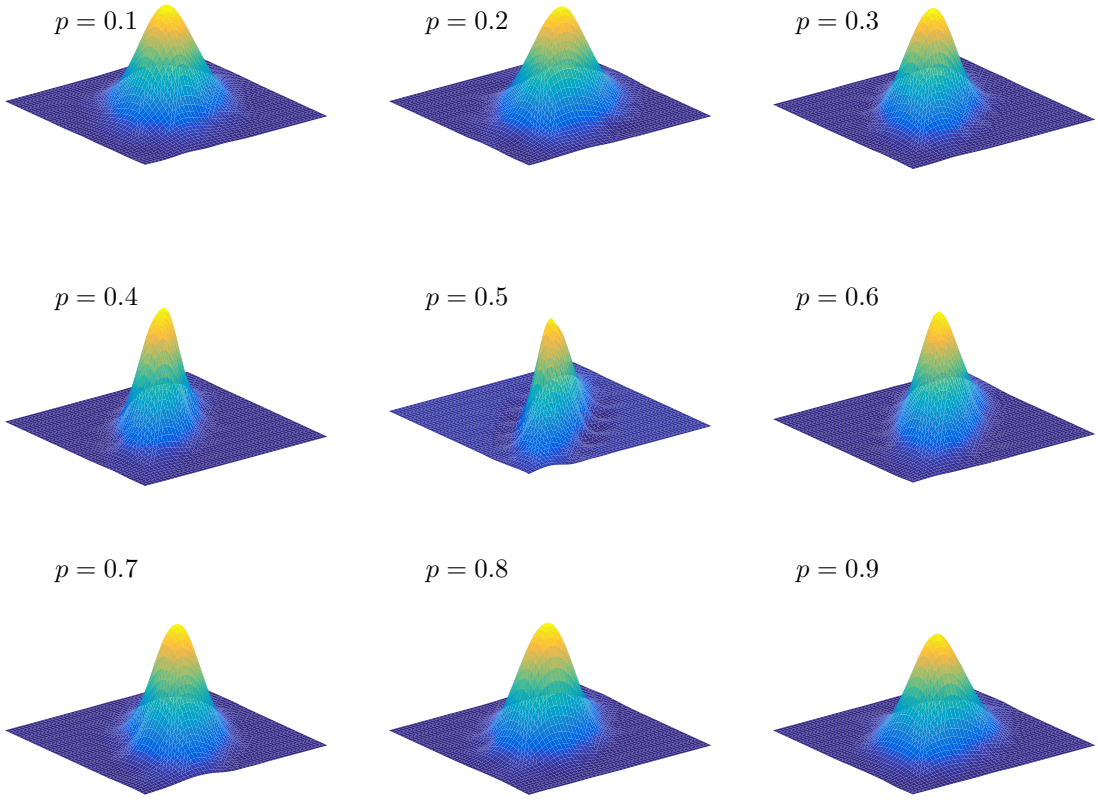

Figure 3: *Joint distributions of* (S_n, K_n) *by Monte-Carlo simulations for* $n = 10^7$ *and varying* p*: the case* $p = 0.5$ *is seen to have stronger dependence than the others.*

trees	$\rho(S_n, K_n)$	$\rho(S_n, N_n)$
tries	$\begin{cases} p \neq q :\to 0 \\ p = q : \text{periodic} \end{cases}$	~ 1
m-ary search trees	$\begin{cases} 3 \leqslant m \leqslant 26 :\to 0 \\ m \geqslant 27 : \text{periodic} \end{cases}$	

Table 2: *A comparison of the correlation coefficients defined on random tries and on random* m*-ary search trees: the size of* m*-ary search trees corresponds to the space requirement, and the KPL and NPL are defined similarly as in tries.*

(or bits generated, or bits inspected) and the number of partitioning rounds in (i) CTM tree algorithm (see Rom and Sidi [25]), (ii) bucket sort (see [18] and Mahmoud et al. [20]), (iii) RS Algorithm for generating random permutations (see Bacher et al. [1]), and (iv) initializing radio networks (see Myoupo et al. [21]). We will also present the results without proof for three other classes of digital trees in the last section.

Our approach is mostly analytic and it is unknown if our results can be characterized by probabilistic arguments. Indeed, we believe that the less expected results we discovered are of

special interest to probabilists as more structural interpretation or characterization remains to be clarified.

An extended abstract of this paper was presented at the 27th International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms (Kraków, Poland; July 4–8, 2016); see [9]. More details of the proofs, as well as a section on extensions are added in this version (some of them in an appendix). Also, we corrected the plots of Figure 3 in [9]. The extended abstract [9] was peer-reviewed and we incorporated the comments and suggestions of the referees into this version.

2 Covariance and Correlation Coefficient

In this section, we prove Theorem A on the asymptotics of the covariance and correlation coefficient of S_n and K_n, where we content ourselves with a detailed sketch of the method because similar proofs have been given in [10]. In fact, we will also need the variances of S_n and K_n, whose derivations will be recalled below and which have been known for some time; see Jacquet and Régnier [14], Kirschenhofer and Prodinger [16], Kirschenhofer et al. [17], Régnier and Jacquet [24]) and [10]. See also Table 1 for a brief summary of these results.

Our method of proof is based on the by-now standard two-stage approach relying on the theory of analytic de-Poissonization and Mellin transform whose origin can be traced back to Jacquet and Régnier [14]. See Flajolet et al. [7] for a survey on Mellin transform, and Jacquet and Szpankowski [15] for a survey on analytic de-Poissonization. For the computation of the covariance, the manipulation can be largely simplified by the additional notions of Poissonized variance and admissible functions further developed in our previous papers [10, 13].

The starting point of our analysis is the recurrence satisfied by S_n and K_n in (1). A standard means in the computation of moments of S_n and K_n is the Poisson generating function, which corresponds to the moments of S_n and K_n with n replaced by a Poisson random variable with parameter z (this step is called *Poissonization*).

More precisely, define the Poisson generating function of $\mathbb{E}(S_n)$ and that of $\mathbb{E}(K_n)$:

$$\tilde{f}_{1,0}(z) := e^{-z}\sum_{n\geqslant 0}\mathbb{E}(S_n)\frac{z^n}{n!}, \quad \text{and} \quad \tilde{f}_{0,1}(z) := e^{-z}\sum_{n\geqslant 0}\mathbb{E}(K_n)\frac{z^n}{n!}.$$

Then the recurrences (1) lead to the functional equations

$$\begin{cases} \tilde{f}_{1,0}(z) = \tilde{f}_{1,0}(pz) + \tilde{f}_{1,0}(qz) + 1 - (1+z)e^{-z}, \\ \tilde{f}_{0,1}(z) = \tilde{f}_{0,1}(pz) + \tilde{f}_{0,1}(qz) + z(1-e^{-z}). \end{cases} \tag{5}$$

From these equations, we obtain, by Mellin transform techniques [7],

$$\tilde{f}_{1,0}(z) \sim z\mathscr{F}[\cdot](z), \qquad \text{and} \qquad \tilde{f}_{0,1}(z) \sim h^{-1}z\log z + z\mathscr{F}[\cdot](z), \tag{6}$$

for large $|z|$ in the half-plane $\Re(z) \geqslant \varepsilon > 0$, where h denotes the entropy of Bernoulli(p). Then, by Cauchy's integral representation and analytic de-Poissonization techniques [15], we obtain precise asymptotic approximations to $\mathbb{E}(S_n)$ and to $\mathbb{E}(K_n)$; see [10] for more details.

Similarly, for the variances $\mathbb{V}(S_n)$ and $\mathbb{V}(K_n)$, we introduce the Poisson generating functions of the second moments:

$$\tilde{f}_{2,0}(z) := e^{-z}\sum_{n\geqslant 0}\mathbb{E}(S_n^2)\frac{z^n}{n!}, \quad \text{and} \quad \tilde{f}_{0,2}(z) := e^{-z}\sum_{n\geqslant 0}\mathbb{E}(K_n^2)\frac{z^n}{n!},$$

which then satisfy, by (1), the same type of functional equations as in (5) but with different non-homogeneous parts. Instead of computing directly asymptotic approximations to the second moments, it proves computational more advantageous to consider the Poissonized variances

$$\begin{cases} \tilde{V}_S(z) := \tilde{f}_{2,0}(z) - \tilde{f}_{1,0}(z)^2 - z\tilde{f}'_{1,0}(z)^2, \\ \tilde{V}_K(z) := \tilde{f}_{0,2}(z) - \tilde{f}_{0,1}(z)^2 - z\tilde{f}'_{0,1}(z)^2, \end{cases} \tag{7}$$

and then following the same Mellin-de-Poissonization approach (as for the means) to derive the first and the third asymptotic estimate in the second column of Table 1; again see [10] for details.

It remains to derive the claimed estimate for the covariance. For that purpose, we introduce the Poisson generating function

$$\tilde{f}_{1,1}(z) := e^{-z} \sum_{n\geqslant 0} \mathbb{E}(S_n K_n) \frac{z^n}{n!},$$

which satisfies, again by (1),

$$\begin{aligned} \tilde{f}_{1,1}(z) = \tilde{f}_{1,1}(pz) + \tilde{f}_{1,1}(qz) + \tilde{f}_{1,0}(pz)\big(\tilde{f}_{0,1}(qz) + z\big) + \tilde{f}_{1,0}(qz)\big(\tilde{f}_{0,1}(pz) + z\big) \\ + pz\tilde{f}'_{1,0}(pz) + qz\tilde{f}'_{1,0}(qz) + \tilde{f}_{0,1}(pz) + \tilde{f}_{0,1}(qz) + z(1 - e^{-z}). \end{aligned}$$

To compute the covariance, it is beneficial to introduce now the *Poissonized covariance* (see (7) or [10] for similar details)

$$\tilde{C}(z) = \tilde{f}_{1,1}(z) - \tilde{f}_{1,0}(z)\tilde{f}_{0,1}(z) - z\tilde{f}'_{1,0}(z)\tilde{f}'_{0,1}(z),$$

which satisfies

$$\tilde{C}(z) = \tilde{C}(pz) + \tilde{C}(qz) + \tilde{h}_1(z) + \tilde{h}_2(z), \tag{8}$$

where

$$\tilde{h}_1(z) = pqz \left(\tilde{f}'_{1,0}(pz) - \tilde{f}'_{1,0}(qz)\right)\left(\tilde{f}'_{0,1}(pz) - \tilde{f}'_{0,1}(qz)\right),$$

and

$$\begin{aligned} \tilde{h}_2(z) = ze^{-z}\left(\tilde{f}_{1,0}(pz) + \tilde{f}_{1,0}(qz) + p(1-z)\tilde{f}'_{1,0}(pz) + q(1-z)\tilde{f}'_{1,0}(qz)\right) \\ + e^{-z}\left((1+z)\tilde{f}_{0,1}(pz) + (1+z)\tilde{f}_{0,1}(qz) - pz^2\tilde{f}'_{0,1}(pz) - qz^2\tilde{f}'_{0,1}(qz)\right) \\ + ze^{-z}\left(1 - (1+z^2)e^{-z}\right). \end{aligned} \tag{9}$$

Note that $\tilde{h}_1$ is zero when $p = \frac{1}{2}$. Furthermore, from (6) (which can be differentiated since they hold in a sector $\mathscr{S} = \{z \in \mathbb{C} \ : \ \Re(z) \geq \epsilon, |\mathrm{Arg}(z)| \leqslant \theta_0\}$ with $0 < \theta_0 < \pi/2$ in the complex plane), we obtain that $\tilde{h}_1(z) = O(|z|)$ and $\tilde{h}_2(z)$ is exponentially small for large $|z|$ in $\Re(z) > 0$. Also $\tilde{h}_1(z) + \tilde{h}_2(z) = O(|z|^2)$ as $z \to 0$. Thus the Mellin transform of $\tilde{h}_1(z) + \tilde{h}_2(z)$ exists in the strip $\langle -2, -1\rangle$, and we have then the inverse Mellin integral representation

$$\tilde{C}(z) = \frac{1}{2\pi i} \int_{-\frac{3}{2}-i\infty}^{-\frac{3}{2}+i\infty} \frac{\mathscr{M}[\tilde{h}_1(z) + \tilde{h}_2(z); s]}{1 - p^{-s} - q^{-s}} z^{-s} \mathrm{d}s, \tag{10}$$

where $\mathscr{M}[\phi(z);s] := \int_0^\infty \phi(z) z^{s-1} \mathrm{d}z$ denotes the Mellin transform of ϕ; see [7].

Next, again from (6) we see that $\mathscr{M}[\tilde{h}_1(z);s]$ can be analytically continued to the vertical line $\Re(s) = -1$ and has no singularities there. Thus, by shifting the line of integration in (10) and computing residues, we obtain

$$\tilde{C}(z) \sim z\mathscr{F}[g^{(2)}](z),$$

uniformly for z in a sector.

What is left is the computation of the Fourier coefficients of the periodic function (see Proposition A below). This is in fact the most technical part of the proof because $\tilde{h}_1(z)$ contains the product of the two terms $\tilde{f}'_{1,0}(pz) - \tilde{f}'_{1,0}(qz)$ and $\tilde{f}'_{0,1}(pz) - \tilde{f}'_{0,1}(qz)$, and thus $\mathscr{M}[\tilde{h}_1(z);s]$ is a Mellin convolution integral. In [10], a general procedure was given for the simplification of such integrals (see [10, p. 24 *et seq.*]). This simplification procedure (see Appendix A for details) and a direct application of the theory of admissible functions of analytic de-Poissonization now yield the following estimate for the covariance of S_n and K_n.

Proposition A. *The covariance of S_n and K_n is asymptotically linear*

$$\mathrm{Cov}(S_n, K_n) \sim n\mathscr{F}[g^{(2)}](n).$$

Here

$$\begin{aligned} g_k^{(2)} = {} & \frac{\Gamma(\chi_k)}{h}\left(1 - \frac{\chi_k+2}{2^{\chi_k+1}}\right) - \frac{1}{h^2}\sum_{j\in\mathbb{Z}\setminus\{0\}} \Gamma(\chi_{k-j}+1)(\chi_j - 1)\Gamma(\chi_j) \\ & - \frac{\Gamma(\chi_k+1)}{h^2}\left(\gamma + 1 + \psi(\chi_k+1) - \frac{p\log^2 p + q\log^2 q}{2h}\right) \\ & + \frac{1}{h}\sum_{\ell\geqslant 2} \frac{(-1)^\ell(p^\ell+q^\ell)}{\ell!(1-p^\ell-q^\ell)}\Gamma(\chi_k+\ell-1)(2\ell^2 - 2\ell + 1 + \chi_k(2\ell-1)), \end{aligned} \tag{11}$$

where γ denotes Euler's constant, $\psi(z)$ is the digamma function and χ_k is defined in (2).

Remark 1. If $\frac{\log p}{\log q} \notin \mathbb{Q}$, then only $k = 0$ is needed and the second term (the sum over j) on the right-hand side of (11) has to be dropped. Also the first term here $\frac{\Gamma(\chi_k)}{h}\left(1 - \frac{\chi_k+2}{2^{\chi_k+1}}\right)$ is taken to be its limit $\frac{1}{h}(\log 2 + \frac{1}{2})$ as $\chi_k \to 0$ when $k = 0$.

The asymptotic estimate for the correlation coefficient in Theorem A now follows from this and the results for the variances of S_n and K_n (see Table 1), where expressions for $g_k^{(1)}$ and $g_k^{(3)}$ can be found, e.g., in [10]. For convenience, we give below the expressions in the unbiased case. Note that both $\mathscr{F}[g^{(1)}](n)$ and $\mathscr{F}[g^{(3)}](n)$ are strictly positive; see Schachinger [26] for details.

In the symmetric case, an alternative expression to (11) (avoiding the convolution of two Fourier series) is

$$g_k^{(2)} = \frac{\Gamma(\chi_k)\left(1 - \frac{\chi_k^2+\chi_k+4}{2^{\chi_k+2}}\right)}{\log 2} + \frac{1}{\log 2}\sum_{\ell\geqslant 1} \frac{(-1)^\ell \Gamma(\chi_k+\ell)\left(\ell(2\ell+1)(\chi_k+\ell) - (\ell+1)^2\right)}{(\ell+1)!(2^\ell-1)};$$

see the discussion of the size of tries in [10], where a similar alternative expression was given for $g_k^{(1)}$, which reads

$$g_k^{(1)} = -\frac{\Gamma(\chi_k-1)\chi_k(\chi_k+1)^2}{4\log 2} + \frac{2}{\log 2}\sum_{\ell\geqslant 1}\frac{(-1)^\ell\Gamma(\chi_k+\ell)\ell\big(\ell(\chi_k+\ell)-1\big)}{(\ell+1)!(2^\ell-1)}.$$

Moreover, also in [10], the following expression for $g_k^{(3)}$ can be found

$$g_k^{(3)} = \frac{\Gamma(\chi_k)\left(1-\frac{\chi_k^2-\chi_k+4}{2^{\chi_k+2}}\right)}{\log 2} + \frac{2}{\log 2}\sum_{\ell\geqslant 1}\frac{(-1)^\ell\Gamma(\chi_k+\ell)(\ell(\chi_k+\ell-1)-1)}{\ell!(2^\ell-1)}.$$

Note that $\chi_k = \frac{2k\pi i}{\log 2}$ and $2^{\chi_k} = 1$, and the reason of retaining 2^{χ_k+2} in the denominator is to give a uniform expression for all k (notably $k = 0$). These provide an explicit expression for the periodic function $F(n)$ in Theorem A. Also, since all the periodic functions have very small amplitude, the average value of the periodic function $F(z)$ can be well-approximated by

$$\frac{g_0^{(2)}}{\sqrt{g_0^{(1)}g_0^{(3)}}} \approx 0.9272416035\cdots.$$

3 Limit Law

In this section, we prove Theorem B, part (i); the proof of part (ii) is similar and only sketched. The key tool of the proof is the multivariate version of the contraction method; see Neininger and Rüschendorf [23]. More precisely, we will use Theorem 3.1 in [23].

We first recall the expression for the square-root of a positive-definite 2×2 matrix

$$M = \begin{pmatrix} a & b\\ b & c\end{pmatrix}.$$

It is well-known that such a matrix has exactly one positive-definite square root which is given by

$$M^{\frac12} = \frac{1}{\sqrt{a+c+2\sqrt{ac-b^2}}}\begin{pmatrix} a+\sqrt{ac-b^2} & b\\ b & c+\sqrt{ac-b^2}\end{pmatrix}, \tag{12}$$

with the inverse

$$M^{-\frac12} = \frac{1}{\sqrt{(ac-b^2)\big(a+c+2\sqrt{ac-b^2}\big)}}\begin{pmatrix} c+\sqrt{ac-b^2} & -b\\ -b & a+\sqrt{ac-b^2}\end{pmatrix}. \tag{13}$$

Now we give the proof of Theorem B, part (i).

Proof of Theorem B, Part (i). Note first that

$$\begin{pmatrix} S_n\\ K_n\end{pmatrix} \overset{d}{=} \begin{pmatrix} 1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix} S_{B_n}\\ K_{B_n}\end{pmatrix} + \begin{pmatrix} 1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix} S^*_{n-B_n}\\ K^*_{n-B_n}\end{pmatrix} + \begin{pmatrix} 1\\ n\end{pmatrix},$$

where the notation is as in Section 1. The contraction method was specially developed for obtaining limiting distribution results for such recurrences; see [23].

We need some notation. First, define

$$\widehat{\Sigma}_n := \begin{pmatrix} \mathbb{V}(S_n) & \mathrm{Cov}(S_n, K_n) \\ \mathrm{Cov}(S_n, K_n) & \mathbb{V}(K_n) \end{pmatrix}. \tag{14}$$

This matrix is clearly positive-definite for all n sufficiently large. Next define

$$M_n^{(1)} := \widehat{\Sigma}_n^{-\frac{1}{2}} \widehat{\Sigma}_{B_n}^{\frac{1}{2}}, \qquad M_n^{(2)} := \widehat{\Sigma}_n^{-\frac{1}{2}} \widehat{\Sigma}_{n-B_n}^{\frac{1}{2}}$$

and

$$\begin{pmatrix} b_n^{(1)} \\ b_n^{(2)} \end{pmatrix} = \widehat{\Sigma}_n^{-\frac{1}{2}} \begin{pmatrix} 1 - \mu(n) + \mu(B_n) + \mu(n - B_n) \\ n - \nu(n) + \nu(B_n) + \nu(n - B_n) \end{pmatrix},$$

where $\mu(n) = \mathbb{E}(S_n)$ and $\nu(n) = \mathbb{E}(K_n)$.

Now to apply the contraction method in [23], it suffices to show that the following conditions hold

$$b_n^{(i)} \xrightarrow{L_3} 0, \qquad M_n^{(i)} \xrightarrow{L_3} M_i, \tag{15}$$

$$\mathbb{E}\big(\|M_1\|_{\mathrm{op}}^3 + \|M_2\|_{\mathrm{op}}^3\big) < 1, \qquad \mathbb{E}\big(\|M_n^{(i)}\|_{\mathrm{op}}^3 \chi_{\{B_n^{(i)} \leqslant j\} \cup \{B_n^{(i)} = n\}}\big) \longrightarrow 0 \tag{16}$$

for $i = 1, 2$ and $j \in \mathbb{N}$, where $\xrightarrow{L_3}$ denotes convergence in the L_3-norm, $\|\cdot\|_{\mathrm{op}}$ is the operator norm, χ_S denotes the characteristic function of set S, $B_n^{(1)} = B_n, B_n^{(2)} = n - B_n$ and

$$M_1 = \begin{pmatrix} \sqrt{p} & 0 \\ 0 & \sqrt{p} \end{pmatrix}, \qquad M_2 = \begin{pmatrix} \sqrt{q} & 0 \\ 0 & \sqrt{q} \end{pmatrix}.$$

Then the contraction method in [23] guarantees that (S_n, K_n) (centralized and normalized) converges in distribution to the unique fixed-point with mean 0, covariance matrix the unity matrix and finite L_3-norm of

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \overset{d}{=} \begin{pmatrix} \sqrt{p} & 0 \\ 0 & \sqrt{p} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} \sqrt{q} & 0 \\ 0 & \sqrt{q} \end{pmatrix} \begin{pmatrix} X_1^* \\ X_2^* \end{pmatrix},$$

where (X_1^*, X_2^*) is an independent copy of (X_1, X_2). Obviously, the bivariate normal distribution is the solution. All this is summarized as follows.

Proposition B. *The following convergence in distribution holds:*

$$\widehat{\Sigma}_n^{-\frac{1}{2}} \begin{pmatrix} S_n - \mathbb{E}(S_n) \\ K_n - \mathbb{E}(K_n) \end{pmatrix} \xrightarrow{d} \mathcal{N}_2(0, I_2).$$

Proof. We only check (15) because the second condition of (16) follows along similar lines and the first condition of (16) follows from (15) in view of

$$\|M_1\|_{\mathrm{op}} = \sqrt{p} \qquad \text{and} \qquad \|M_2\|_{\mathrm{op}} = \sqrt{q}.$$

We start with proving (15) for $b_n^{(i)}$ for which we use the notations

$$\Omega_1(n) = \mathbb{V}(S_n), \quad \Omega_2(n) = \mathrm{Cov}(S_n, K_n), \quad \Omega_3(n) = \mathbb{V}(K_n)$$

and

$$D(n) = \Omega_1(n)\Omega_3(n) - \Omega_2(n)^2.$$

Also define

$$R(n) = \Omega_1(n) + \Omega_3(n) + 2\sqrt{D(n)}.$$

Then, by (12), we see that

$$b_n^{(1)} = (1 - \mu(n) + \mu(B_n) + \mu(n - B_n))\frac{\Omega_3(n) + \sqrt{D(n)}}{\sqrt{D(n)R(n)}} - (n - \nu(n) + \nu(B_n) + \nu(n - B_n))\frac{\Omega_2(n)}{\sqrt{D(n)R(n)}}$$

and a similar expression for $b_n^{(2)}$ holds. From the normality of both S_n and K_n (proved for S_n via the contraction method in [11] and a similar method of proof also applies to K_n), we have

$$\frac{1 - \mu(n) + \mu(B_n) + \mu(n - B_n)}{\sqrt{n}} \xrightarrow{L_3} 0$$

and

$$\frac{n - \nu(n) + \nu(B_n) + \nu(n - B_n)}{\sqrt{n \log n}} \xrightarrow{L_3} 0.$$

Moreover, we have

$$\sqrt{n}\,\frac{\Omega_3(n) + \sqrt{D(n)}}{\sqrt{D(n)R(n)}} \sim \frac{1}{\sqrt{\mathscr{F}[g^{(1)}](n)}},$$

and

$$\sqrt{n \log n}\,\frac{\Omega_2(n)}{\sqrt{D(n)R(n)}} \sim \frac{\mathscr{F}[g^{(2)}](n)}{\lambda\sqrt{\log n \mathscr{F}[g^{(1)}](n)}},$$

where $g^{(1)}, g^{(2)}$ and λ are as above. Thus, both sequences are bounded and, consequently, we obtain the claimed result with L_3-convergence above. Similarly, one proves (15) for $b_n^{(2)}$.

Next, we consider $M_n^{(i)}$. Here, we only show the claim for the $(1,1)$ entry of $M_n^{(1)}$ (denoted by $M_n^{(1)}(1,1)$) all other cases being treated similarly. First, observe that by definition and matrix square-root, we have

$$M_n^{(1)}(1,1) = \frac{\sqrt{R(n)}}{\sqrt{R(B_n)}} \cdot \frac{(\Omega_3(n) + \sqrt{D(n)})(\Omega_1(B_n) + \sqrt{D(B_n)}) - \Omega_2(n)\Omega_2(B_n)}{\sqrt{D(n)R(n)}}.$$

Now, from the strong law of large numbers for the binomial distribution

$$\frac{B_n}{n} \xrightarrow{\text{a.s.}} p$$

and from Taylor series expansion (note that all periodic functions are infinitely differentiable), we have

$$\frac{\sqrt{R(n)}}{\sqrt{R(B_n)}} \xrightarrow{\text{a.s.}} \frac{1}{\sqrt{p}},$$

and

$$\frac{(\Omega_3(n)+\sqrt{D(n)})(\Omega_1(B_n)+\sqrt{D(B_n)})-\Omega_2(n)\Omega_2(B_n)}{\sqrt{D(n)R(n)}} \xrightarrow{\text{a.s.}} p.$$

Thus, $M_n^{(1)}(1,1) \xrightarrow{\text{a.s.}} \sqrt{p}$ from which the claim follows by the dominated convergence theorem. ∎

Next, set

$$\widetilde{\Sigma}_n := \begin{pmatrix} n\mathscr{F}[g^{(1)}](n) & 0 \\ 0 & \lambda n \log n \end{pmatrix}.$$

Then, we have the following simple lemma.

Lemma 1. *We have, as $n \to \infty$,*

$$\widehat{\Sigma}_n^{-\frac{1}{2}} \widetilde{\Sigma}_n^{\frac{1}{2}} \to I_2.$$

Proof. This follows by a straightforward computation using the expressions for the matrix square-root (12) and its inverse (13). For example, the entry $(1,2)$ of $\widehat{\Sigma}_n^{-\frac{1}{2}} \widetilde{\Sigma}_n^{\frac{1}{2}}$ (where we use the notations from the proof of the previous proposition) satisfies

$$-\frac{\Omega_2(n)\sqrt{\lambda n \log n}}{\sqrt{D(n)R(n)}} \sim -\frac{\mathscr{F}[g^{(2)}](n)}{\sqrt{\lambda \log n \mathscr{F}[g^{(1)}](n)}},$$

which tends to 0 as claimed. The other entries are treated similarly, ∎

Theorem B, part (i) now follows from this lemma and Proposition B.

Next, we sketch the (similar) proof of Theorem B, part (ii).

Proof of Theorem B, Part (ii). The proof runs along similar lines as in Part (i). The only difference is that now it is not entirely obvious that $\widehat{\Sigma}_n$ is positive definite. Note, however, that from the discussion in the introduction, this matrix is positive-definite if and only if Σ_n (defined in Theorem B) is positive definite. This is ensured by the following lemma.

Lemma 2. *Σ_n is positive-definite for all n large enough.*

Proof. It suffices to show that $\det(\Sigma_n) > 0$ for all n large enough. Indeed, we have

$$\begin{aligned} \det(\Sigma_n) &= n^2 \mathscr{F}[g^{(1)}](n)\mathscr{F}[g^{(3)}](n) - n^2(\mathscr{F}[g^{(1)}](n))^2 \\ &= n^2 \mathscr{F}[g^{(1)}](n)\mathscr{F}[g^{(3)}](n)(1 - F(\log_2 n)^2), \end{aligned}$$

from which the result follows.

Note that this in addition shows the stronger result $\det(\Sigma_n) \geqslant dn^2$ for all n large enough where $d > 0$. (A proof avoiding numerical computations can be performed using the same approach as in Proposition 3 of [12].) ∎

The rest of the proof is similar as in the asymmetric case and is omitted.

4 Extensions

In this section, we show that the dependence phenomena we discovered here on random binary tries (Theorem A and Theorem B) also find their appearance in other trees and structures whose subtree-sizes and sub-structure-sizes are dictated by a binomial or a multinomial distribution.

For simplicity, we consider in this section only three varieties of random digital trees: random m-ary tries, random PATRICIA tries and random bucket digital search trees; see [10] for more potential examples with the same splitting principles.

m-ary Tries. It is straightforward to extend our tries constructed from binary input strings to inputs from an m-ary alphabets, $m \geqslant 2$. In this case, the resulting trie becomes an m-ary tree (since each node now has m subtrees one belonging to each letter). As a random model, we assume that bits are generated independently at random with the i-th letter occurring with probability p_i, where $p_1 + \cdots + p_m = 1$ and $0 < p_i < 1$ for $1 \leqslant i \leqslant m$.

The size and the key path length (which we again denote by S_n and K_n) in such random m-ary tries satisfy the recurrences

$$\begin{cases} S_n \stackrel{d}{=} S^{(1)}_{I^{(1)}_n} + \cdots + S^{(m)}_{I^{(m)}_n} + 1, \\ K_n \stackrel{d}{=} K^{(1)}_{I^{(1)}_n} + \cdots + K^{(m)}_{I^{(m)}_n} + n, \end{cases} \qquad (n \geqslant 2),$$

with the initial conditions $S_n = K_n = 0$ for $n \leqslant 1$, where $(S^{(i)}_n)$ and $(K^{(i)}_n)$ are independent copies of (S_n) and (K_n), respectively, for $1 \leqslant i \leqslant m$, and

$$\mathbb{P}(I^{(1)}_n = j_1, \ldots, I^{(m)}_n = j_m) = \binom{n}{j_1, \ldots, j_m} p_1^{j_1} \cdots p_m^{j_m}, \tag{17}$$

for all $j_1, \ldots, j_m \geqslant 0$ with $j_1 + \cdots + j_m = n$.

The pair (S_n, K_n) satisfies the same type of properties as those described in Theorem A and Theorem B for binary tries, where the symmetric case here corresponds to $p_1 = \cdots = p_m = 1/m$ and all other cases are asymmetric. Only the expressions for $g^{(1)}_k, g^{(2)}_k, g^{(3)}_k$ and λ are different but they can be computed via the same analytic tools as those used in [10]. For the sake of simplicity, we only give the expressions in the symmetric case ($\chi_k = 2k\pi i/\log m$) as follows:

$$g^{(1)}_k = \frac{\Gamma(\chi_k - 1)\big(\chi_k - \frac{\chi_k^3 + 2\chi_k^2 + 5\chi_k}{2^{\chi_k+2}}\big)}{\log m} + \frac{2}{\log m} \sum_{\ell \geqslant 1} \frac{(-1)^\ell \Gamma(\chi_k + \ell)\ell\big(\ell(\chi_k + \ell) - 1\big)}{(\ell+1)!(m^\ell - 1)},$$

$$g^{(2)}_k = \frac{\Gamma(\chi_k)\big(1 - \frac{\chi_k^2 + \chi_k + 4}{2^{\chi_k+2}}\big)}{\log m} + \frac{1}{\log m} \sum_{\ell \geqslant 1} \frac{(-1)^\ell \Gamma(\chi_k + \ell)\,\big(\ell(2\ell+1)(\chi_k + \ell) - (\ell+1)^2\big)}{(\ell+1)!(m^\ell - 1)},$$

$$g^{(3)}_k = \frac{\Gamma(\chi_k)\big(1 - \frac{\chi_k^2 - \chi_k + 4}{2^{\chi_k+2}}\big)}{\log m} + \frac{2}{\log m} \sum_{\ell \geqslant 1} \frac{(-1)^\ell \Gamma(\chi_k + \ell)(\ell(\chi_k + \ell - 1) - 1)}{\ell!(m^\ell - 1)}.$$

Note that the variance of the size was considered in [12], but no explicit expression was given for the Fourier coefficients of the periodic function.

With the help of these expressions, we obtain the following numerical approximations to the average value of the periodic function of the correlation coefficient between S_n and K_n in the symmetric case. We see that they differ little.

PATRICIA Tries. A simple idea to increase the efficiency of tries is to remove all internal nodes with one-way branching. The resulting tree is called a PATRICIA trie; here PATRICIA is an acronym of "Practical Algorithm To Retrieve Information Coded In Alphanumeric".

m	2	3	4	5	6
average value of the periodic fluctuations	0.927	0.925	0.924	0.922	0.921

Table 3: *Numerical approximations to the average values of the periodic functions $F(n)$ arising in the asymptotic estimate $\rho(S_n, K_n) \sim F(n)$ for the symmetric case and $m = 2, \dots, 6$.*

We use the same random model as we used above for m-ary tries and consider the size and key-path length of PATRICIA tries (which we again denote by S_n and K_n). Then they satisfy the recurrences

$$\begin{aligned} S_n &\stackrel{d}{=} S^{(1)}_{I^{(1)}_n} + \cdots + S^{(m)}_{I^{(m)}_n} + T_n, \\ K_n &\stackrel{d}{=} K^{(1)}_{I^{(1)}_n} + \cdots + K^{(m)}_{I^{(m)}_n} + nT_n, \end{aligned} \qquad (n \geqslant 2),$$

with the initial conditions $S_n = K_n = 0$ for $n \geqslant 1$, where $I^{(i)}_n$ is defined as in (17) above, $(S^{(i)}_n)$ and $(K^{(i)}_n)$ are independent copies of (S_n) and (K_n), respectively, and

$$T_n = \begin{cases} 1, & \text{if } I^{(i)}_n < n \text{ for all } 1 \leqslant i \leqslant m; \\ 0, & \text{otherwise.} \end{cases}$$

Note that for $m = 2$, the size is deterministic. We thus assume $m \geqslant 3$ to avoid trivialities. Then the dependence of (S_n, K_n) satisfies *mutatis mutandis* Theorem A and Theorem B. In particular, the required changes for $g^{(1)}_k, g^{(2)}_k, g^{(3)}_k$ in the symmetric case are given as follows ($\chi_k = 2k\pi i/\log m$):

$$\begin{aligned} g^{(1)}_k &= \frac{(m-1)\Gamma(\chi_k - 1)}{\log m}\Bigg(-1 - \frac{(m-1)(\chi_k+1)}{2\chi_k} \\ &\quad + \left(1-\frac{1}{m}\right)^{-\chi_k}\left(1 - \frac{(m-1)\chi_k + m + 1}{2\chi_k}\right) + \left(2 - \frac{1}{m}\right)^{-\chi_k}(2(m-1)\chi_k + 2m)\Bigg) \\ &\quad + \frac{2(m-1)^2}{\log m}\sum_{\ell\geqslant 1}\frac{(-1)^{\ell+1}\Gamma(\ell+\chi_k)\ell\left(1-\left(1-\frac{1}{m}\right)^{\ell}\right)\left(1-\left(1-\frac{1}{m}\right)^{-\ell-\chi_k}\right)}{(\ell+1)!(m^\ell-1)}, \\ g^{(2)}_k &= \frac{\Gamma(\chi_k)}{\log m}\left(\left(1-\frac{1}{m}\right)^{-\chi_k}\left(1 - \frac{(m-1)\chi_k + 2}{2\chi_{k+1}}\right) + \left(2-\frac{1}{m}\right)^{-\chi_k}\frac{(m-1)^2\chi_k}{2m-1}\right) \\ &\quad + \frac{m-1}{\log m}\sum_{\ell\geqslant 1}\frac{(-1)^{\ell+1}\Gamma(\ell+\chi_k)\ell}{(\ell+1)!(m^\ell-1)}\Bigg((\ell+1)\left(1-\frac{1}{m}\right)^{\ell} \\ &\quad + (\chi_k - 1)\left(1-\frac{1}{m}\right)^{-\chi_k} - (\ell+\chi_k)\left(1-\frac{1}{m}\right)^{-\ell-\chi_k}\Bigg), \\ g^{(3)}_k &= \frac{\Gamma(\chi_k)}{\log m}\left(1-\frac{1}{m}\right)^{-\chi_k}\left(1 + \frac{\chi_k}{m-1} - \frac{(m-1)\chi_k^2 - (m-3)\chi_k + 4(m-1)}{(m-1)2^{\chi_k+2}}\right) \\ &\quad + \frac{2(m-1)^{-\chi_k}}{\log m}\sum_{\ell\geqslant 1}\frac{(-1)^{\ell}\Gamma(\ell+\chi_k+1)}{(\ell-1)!(m^\ell-1)}. \end{aligned}$$

These expressions are also valid for $m = 2$, where $g_k^{(1)}$ and $g_k^{(2)}$ can be shown to be identically zero. Note that the result for the variance of the key-path length was already derived in [10] (for $m = 2$) and that for the size was established in [12] but without a precise expression for the Fourier coefficients.

Again, we can use the above expressions to obtain the average value of the periodic function of the correlation coefficient between S_n and K_n in the symmetric case. Note that unlike tries, these values increase with m.

m	3	4	5	6
average value of the periodic fluctuations	0.751	0.814	0.841	0.856

Table 4: *Numerical approximations to the average values of the periodic functions $F(n)$ arising in the asymptotic estimate $\rho(S_n, K_n) \sim F(n)$ for the symmetric case and $m = 3, \dots, 6$.*

Bucket Digital Search Trees. Digital search trees (DST) represent yet another class of digital tree structures; see [18, 19] for more information. In contrast to tries and PATRICIA tries, they only have one type of nodes where data are stored. More precisely, given a set of data consisting of n infinite 0-1 strings, a DST is constructed as follows: if $n = 1$, then the DST consists of only one node holding the sole string; otherwise, the first string is stored in the root and all others are directed to the subtrees according to their first bit being 0 or 1; then, the subtrees are built recursively but by using consecutive bits to split the data.

Clearly, the size of such a DST is deterministic and equals the input cardinality. We consider instead a bucket version with an additional capacity $b \geqslant 2$, allowing each node holding up to b strings and nodes having subtrees only when they are filled up.

We adopt the same Bernoulli random model as for random tries and consider the size and key-path length in random bucket digital search trees (again denoted by S_n and K_n), which then satisfy

$$\begin{cases} S_{n+b} \stackrel{d}{=} S_{B_n} + S^*_{n-B_n} + 1, \\ K_{n+b} \stackrel{d}{=} K_{B_n} + K^*_{n-B_n} + n, \end{cases} \qquad (n \geqslant 0),$$

with the initial conditions $S_0 = K_0 = K_1 = \cdots = K_{b-1} = 0$ and $S_1 = \cdots = S_{b-1} = 1$, where (S^*_n) and (K^*_n) are independent copies of (S_n) and (K_n), respectively.

The same dependence phenomena as those described in Theorem A and Theorem B also hold for the pair (S_n, K_n). The computation of the sequences $g_k^{(1)}, g_k^{(2)}, g_k^{(3)}$ is nevertheless more intricate. In the asymmetric case, one can again use analytic de-Poissonization and Mellin transform techniques, however, the resulting expressions are less explicit. On the other hand, in the symmetric case, explicit expressions for $g_k^{(1)}, g_k^{(2)}, g_k^{(3)}$ are available via the Poisson-Laplace-Mellin method from [13]. As the expressions are long, we omit them here. Note that the results for the variances of S_n and K_n have already been obtained in [13].

Other Shape Parameters. Theorem A and Theorem B also extend to pairs of random variables where the size is replaced by the number of various patterns (such as the number of internal-external nodes discussed, e.g., by Flajolet and Sedgewick in [8]) and the key-path

length is replaced by other notions of the path length (such as the total path length of internal-external nodes).

Acknowledgments

The first author acknowledges partial supported by MOST under the grants MOST-104-2923-M-009-006-MY3 and MOST-105-2115-M-009-010-MY2. We also thank the helpful comments by the referees for the extended abstract [9] of this paper.

References

[1] A. Bacher, O. Bodini, H.-K. Hwang and T.-H. Tsai (2016). Generating random permutations by coin-tossing: classical algorithms, new analysis and modern implementation, *ACM Trans. Algorithms*, accepted for publication.

[2] H.-H. Chern, M. Fuchs, H.-K. Hwang and R. Neininger (2016). Dependencies and phase changes in random m-ary search trees, *Random Struct. Algor.*, accepted for publication.

[3] J. Clément, P. Flajolet and B. Vallée (1998). Dynamical sources in information theory: a general analysis of trie structures, *Algorithmica*, **29**, 307–369.

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

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

[6] P. Flajolet (2006). The ubiquitous digital tree. In *Lecture Notes in Comput. Sci.* (*STACS 2006*), **3884**, pp. 1–22, Springer, Berlin.

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

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

[9] M. Fuchs and H.-K. Hwang (2016). Dependence between external path-length and size in random tries, 27th International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms, Paper **10**.

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

[11] M. Fuchs and C.-K. Lee (2014). A general central limit theorem for shape parameters of m-ary tries and PATRICIA tries, *Electron. J. Combin.*, **21**, 26 pages.

[12] M. Fuchs and C.-K. Lee (2015). The Wiener index of random digital trees, *SIAM J. Discrete Math.*, **29**, 586–614.

[13] H.-K. Hwang, M. Fuchs and V. Zacharovas (2010). Asymptotic variance of random symmetric digital search trees, *Discrete Math. Theor. Comput. Sci.*, **12**, 103–166.

[14] P. Jacquet and M. Régnier (1986). Trie partitioning process: limiting distributions. In *CAAP 86* (Nice, 1986), vol. **214** of *Lecture Notes in Comput. Sci.*, Springer, Berlin, 196–210.

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

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

[17] P. Kirschenhofer, H. Prodinger and W. Szpankowski (1989). On the variance of the external path length in a symmetric digital trie, *Discrete Appl. Math.*, **25**, 129–143.

[18] D. E. Knuth (1998). *The Art of Computer Programming, Volume 3. Sorting and Searching*, Second edition, Addison-Wesley, Reading, MA.

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

[20] H. M. Mahmoud, P. Flajolet, P. Jacquet and M. Régnier (2000). Analytic variations on bucket selection and sorting, *Acta Inform.*, **36**, 735–760.

[21] J.-F. Myoupo, L. Thimonier and V. Ravelomanana (2003). Average case analysis-based protocols to initialize packet radio networks, *Wireless Comm. Mobile Comput.*, **3**, 539–548.

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

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

[24] M. Régnier and P. Jacquet (1989). New results on the size of tries, *IEEE Trans. Inform. Theory*, **35**, 203–205.

[25] R. Rom and M. Sidi (1990). *Multiple Access Protocols: Performance and Analysis*, Springer Science & Business Media.

[26] W. Schachinger (1995), On the variance of a class of inductive valuations of data structures for digital search, *Theoret. Comput. Sci.*, **144**, 251–275.

[27] Y. L. Tong (1990). *The Multivariate Normal Distribution*, Springer-Verlag, New York.

A A Sketch of Proof of (11)

We sketch here some details of how the expression (11) for $g_k^{(2)}$ in Proposition A is obtained. The method we use is based on that introduced in [10]; see p. 25 *et seq.*

First, by moving the line of integration in (10) to the right and using the residue theorem, we have

$$g_k^{(2)} = \frac{G_2(-1+\chi_k)}{h},$$

where $G_2(s) = \mathscr{M}[\tilde{h}_1(z) + \tilde{h}_2(z); s]$. Note that in the above expression and in what follows, if $\log p/\log q$ is irrational, then only the term with $k = 0$ is retained. Thus, our problem boils down to the computation of $G_2(s)$.

We first consider the Mellin transform of $\tilde{h}_2(z)$ which is easier to handle. By the expression (9) for $\tilde{h}_2(z)$ from Section 2, the Mellin transform is given by

$$\mathscr{M}[\tilde{h}_2(z); s] = \Gamma(s+1)\left(1 - \frac{s^2+3s+6}{2^{s+3}}\right) + Y(s),$$

where

$$\begin{aligned} Y(s) = \int_0^\infty e^{-z}\Big(&z\tilde{f}_{1,0}(pz) + pz(1-z)\tilde{f}'_{1,0}(pz) + z\tilde{f}_{1,0}(qz) + qz(1-z)\tilde{f}'_{1,0}(qz) \\ &+ (1+z)\tilde{f}_{0,1}(pz) - pz^2\tilde{f}'_{0,1}(pz) + (1+z)\tilde{f}_{0,1}(qz) - qz^2\tilde{f}'_{0,1}(qz)\Big)z^{s-1}\mathrm{d}z. \end{aligned}$$

Observe that by applying the Mellin transform and its inverse to (5), we obtain

$$\tilde{f}_{1,0}(z) = \frac{1}{2\pi i}\int_{(-3/2)} \frac{-(\omega+1)\Gamma(\omega)}{1-p^{-\omega}-q^{-\omega}}\, z^{-\omega}\mathrm{d}\omega \tag{18}$$

and

$$\tilde{f}_{0,1}(z) = \frac{1}{2\pi i}\int_{(-3/2)} \frac{-\omega\Gamma(\omega)}{1-p^{-\omega}-q^{-\omega}}\, z^{-\omega}\mathrm{d}\omega. \tag{19}$$

Substituting these into the integral representation of $Y(s)$ and interchanging the integrals, we see that

$$\begin{aligned} Y(s) &= \frac{1}{2\pi i}\int_{(-3/2)} \frac{\Gamma(\omega)\Gamma(s-\omega)}{1-p^{-\omega}-q^{-\omega}}\left(p^{-\omega}+q^{-\omega}\right)\left(\omega^2-(s-\omega)(2\omega^2+3\omega+1)\right)\mathrm{d}\omega \\ &= \sum_{\ell\geqslant 2}\frac{(-1)^\ell(p^\ell+q^\ell)}{\ell!(1-p^\ell-q^\ell)}\Gamma(s+\ell)\left(\ell^2-(s+\ell)(2\ell^2-3\ell+1)\right), \end{aligned}$$

where the last line follows from moving the vertical line of integration to minus infinity and summing over all the residues of the poles encountered.

For $\mathscr{M}[\tilde{h}_1(z); -1+\chi_k]$, we use the expression for $\tilde{h}_1(z)$ in Section 2, (18) and (19), and Mellin convolution, giving

$$\frac{1}{2\pi i}\int_{(0)+} \frac{pq\left(p^{-\omega}-q^{-\omega}\right)\left(p^{\omega}-q^{\omega}\right)}{(1-p^{1-\omega}-q^{1-\omega})(1-p^{1+\omega}-q^{1+\omega})}\Gamma(\omega+1)(\chi_k-\omega-1)\Gamma(\chi_k-\omega)\mathrm{d}\omega,$$

where the integration path is the imaginary axis with a small indentation to the right at the zeros of $1-p^{1-\omega}-q^{1-\omega}$. Now by the decomposition

$$\frac{pq\left(p^{-\omega}-q^{-\omega}\right)\left(p^{\omega}-q^{\omega}\right)}{(1-p^{1-\omega}-q^{1-\omega})(1-p^{1+\omega}-q^{1+\omega})} = \frac{1}{1-p^{1-\omega}-q^{1-\omega}} + \frac{p^{1+\omega}+q^{1+\omega}}{1-p^{1+\omega}-q^{1+\omega}},$$

the above integral is rewritten as

$$\frac{1}{2\pi i}\int_{(0)+}\left(\frac{1}{1-p^{1-\omega}-q^{1-\omega}}+\frac{p^{1+\omega}+q^{1+\omega}}{1-p^{1+\omega}-q^{1+\omega}}\right)\Gamma(\omega+1)(\chi_k-\omega-1)\Gamma(\chi_k-\omega)\mathrm{d}\omega. \quad (20)$$

We break now this integral into two parts according to the two terms in the bracket. For the first part, we use the substitution $\omega \leftrightarrow \chi_k - \omega$ and standard residue calculus, and obtain

$$\begin{aligned}
&\frac{1}{2\pi i}\int_{(0)-}\frac{1}{1-p^{1+\omega}-q^{1+\omega}}\,\Gamma(\chi_k-\omega+1)(\omega-1)\Gamma(\omega)\mathrm{d}\omega\\
&\quad=-\frac{\Gamma(\chi_k+1)}{h}\left(\gamma+1+\psi(\chi_k+1)-\frac{p\log^2 p+q\log^2 q}{2h}\right)\\
&\qquad-\frac{1}{h}\sum_{j\in\mathbb{Z}\setminus\{0\}}\Gamma(\chi_{k-j}+1)(\chi_j-1)\Gamma(\chi_j)\\
&\qquad+\frac{1}{2\pi i}\int_{(0)+}\frac{1}{1-p^{1+\omega}-q^{1+\omega}}\,\Gamma(\chi_k-\omega+1)(\omega-1)\Gamma(\omega)\mathrm{d}\omega,
\end{aligned}$$

where the second line follows by moving the line of integration over the imaginary axis and $\psi(s)$ denotes the derivative of $\log\Gamma(s)$. Next, note that

$$\begin{aligned}
&\frac{1}{2\pi i}\int_{(0)+}\frac{1}{1-p^{1+\omega}-q^{1+\omega}}\,\Gamma(\chi_k-\omega+1)(\omega-1)\Gamma(\omega)\mathrm{d}\omega\\
&\quad=\frac{1}{2\pi i}\int_{(0)+}\Gamma(\chi_k-\omega+1)(\omega-1)\Gamma(\omega)\mathrm{d}\omega\\
&\qquad+\frac{1}{2\pi i}\int_{(0)+}\frac{p^{1+\omega}+q^{1+\omega}}{1-p^{1+\omega}-q^{1+\omega}}\Gamma(\chi_k-\omega+1)(\omega-1)\Gamma(\omega)\mathrm{d}\omega.
\end{aligned}$$

The first integral on the right-hand side is a Mellin convolution integral and can be evaluated explicitly as

$$\begin{aligned}
\frac{1}{2\pi i}\int_{(0)+}\Gamma(\chi_k-\omega+1)(\omega-1)\Gamma(\omega)\mathrm{d}\omega&=\int_0^\infty e^{-z}z^{\chi_k}(1-(1-z)e^{-z})\mathrm{d}z-\Gamma(\chi_k+1)\\
&=\Gamma(\chi_k+1)\frac{\chi_k-1}{2^{\chi_k+2}}.
\end{aligned}$$

For the second integral, we move the line of integration to infinity and use the residue theorem, yielding

$$\begin{aligned}
&\frac{1}{2\pi i}\int_{(0)+}\frac{p^{1+\omega}+q^{1+\omega}}{1-p^{1+\omega}-q^{1+\omega}}\,\Gamma(\chi_k-\omega+1)(\omega-1)\Gamma(\omega)\mathrm{d}\omega\\
&\quad=\sum_{\ell\geqslant 2}\frac{(-1)^\ell(p^\ell+q^\ell)}{(\ell-1)!(1-p^\ell-q^\ell)}\,\Gamma(\chi_k+\ell-1)(\ell-1)(\chi_k+\ell-2).
\end{aligned}$$

In a similar way, the second part of (20) has the series representation

$$\begin{aligned}
&\frac{1}{2\pi i}\int_{(0)+}\frac{p^{1+\omega}+q^{1+\omega}}{1-p^{1+\omega}-q^{1+\omega}}\,\Gamma(\omega+1)(\chi_k-\omega-1)\Gamma(\chi_k-\omega)\mathrm{d}\omega\\
&\quad=\sum_{\ell\geqslant 2}\frac{(-1)^\ell(p^\ell+q^\ell)}{(\ell-1)!(1-p^\ell-q^\ell)}\,\Gamma(\chi_k+\ell-1)\ell(\chi_k+\ell-1).
\end{aligned}$$

Since

$$\begin{aligned} G_2(-1+\chi_k) &= \mathscr{M}[\tilde{h}_1(z);-1+\chi_k] + \mathscr{M}[\tilde{h}_2(z);-1+\chi_k] \\ &= \mathscr{M}[\tilde{h}_1(z);-1+\chi_k] + \Gamma(\chi_k)\left(1-\frac{\chi_k^2+\chi_k+4}{2^{\chi_k+2}}\right) + Y(-1+\chi_k), \end{aligned}$$

we then deduce (11) by collecting all expressions.