Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Dependence between External Path-Length and Size in Random Tries

M. Fuchs[1]†and H.-K. Hwang[2]

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

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 that the internal path length is totally positively correlated to the size and that both tend to the same normal limit law. These two examples provide concrete instances of bivariate normal distributions (as limit laws) whose correlation is 0, 1 and periodically oscillating.

Keywords: Random tries, Pearson's correlation coefficient, asymptotic normality, Poissonization/de-Poissonization, Mellin transform, contraction method

1 Introduction

Tries are one of the most fundamental tree-type data structures in computer algorithms. 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 IPL and the size are known to have asymptotic correlation coefficient tending to one and to have the same normal limit law after each being properly normalized (see [4, 6]), this paper aims to show that the EPL exhibits a completely different behavior depending on the parameter of the underlying random bits being biased or unbiased. This is a companion paper to [1].

Given a sequence of binary strings (or keys), one can construct a (binary) trie as follows. If $n = 1$, then the trie consists of a single root-node holding the sole string; otherwise, 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 going to the same subtree are then constructed 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.

†Partially supported by MOST under the grant MOST-104-2923-M-009-006-MY3

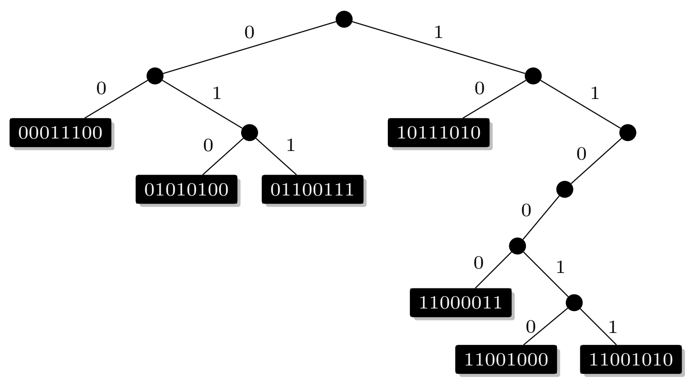

Fig. 1: *A trie with $n = 7$ records: the (filled) circles represent internal nodes and rectangles holding the binary strings are external nodes. In this example, $S_n = 8$, $K_n = 27$, and $N_n = 18$.*

The random trie model we consider here assumes that each of the n binary keys is an infinite sequence consisting 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;
- IPL (or node path-length, NPL) N_n: the sum of the distances between the root to each internal node;
- EPL (or key path-length, KPL) K_n: the sum of the distances between the root to 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 in random m-ary search trees in the companion paper [1]; see below for more details.

By the recursive definition, we have the following recurrence relations

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

where $B_n = \text{Binom}(n, p)$ and $S_0 = S_1 = K_0 = K_1 = N_0 = N_1 = 0$. Here $(S^*_n), (K^*_n)$, 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 [4] and the references therein), 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 [4], 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 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 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 of the above three shape parameters are summarized in the following table; see [4] 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	$\mathbb{E}(D_n) = \frac{\mathbb{E}(K_n)}{n}$	$\mathbb{V}(D_n) = \frac{\mathbb{V}(K_n)}{n} + O(1)$

Tab. 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. Asymptotic normality holds for all three random variables S_n, N_n, K_n.*

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 the dependence and independence of K_n on S_n.*

On the other hand, if we denote by D_n the depth, which is defined to be the distance between the root and a randomly chosen external node (each with the same probability), then we have not only the relation $\mathbb{E}(D_n)n = \mathbb{E}(K_n)$, but also the asymptotic equivalent $\mathbb{V}(D_n)n \sim \mathbb{V}(K_n)$ when $p \neq 1/2$ (or $\lambda > 0$), and a central limit theorem holds; see Devroye [2].

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 [4] 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 [6], 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 [16]), we can write

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

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.*

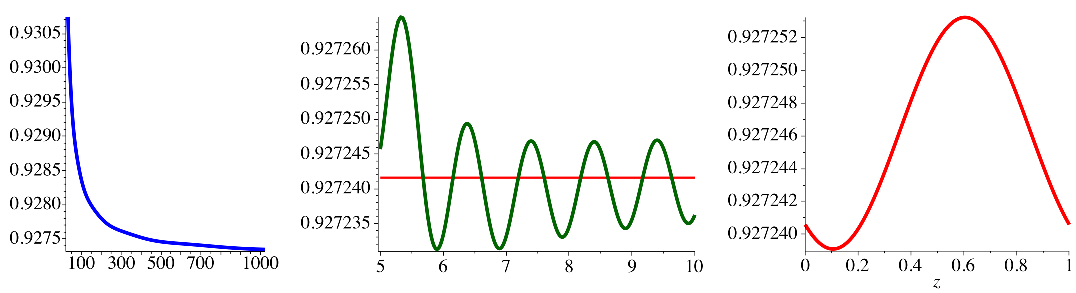

Fig. 2: $p = \frac{1}{2}$*: periodic fluctuations of (i) $\rho(S_n, K_n)$ (left) for $n = 32, \ldots, 1024$, (ii) $\frac{\mathrm{Cov}(S_n,K_n)}{\sqrt{\mathbb{V}(S_n)(\mathbb{V}(K_n)+1.046)}}$ (middle) in logarithmic scale, and (iii) $F(n)$ by its Fourier series expansion (right). Note that the fluctuations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

Furthermore, we show that this different behavior cannot be ascribed to the weak measurability of nonlinear dependence of Pearson's correlation coefficient ρ since the same dependence is also present in the limiting distribution. (For the univariate central limit theorems implied by the result below, see Jacquet and Régnier [8] 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.

(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 [16]), it is more transparent to split the statement into two cases. See Figure 3 for 3D-plots of the joint distributions of (S_n, K_n) when $n = 10^7$.

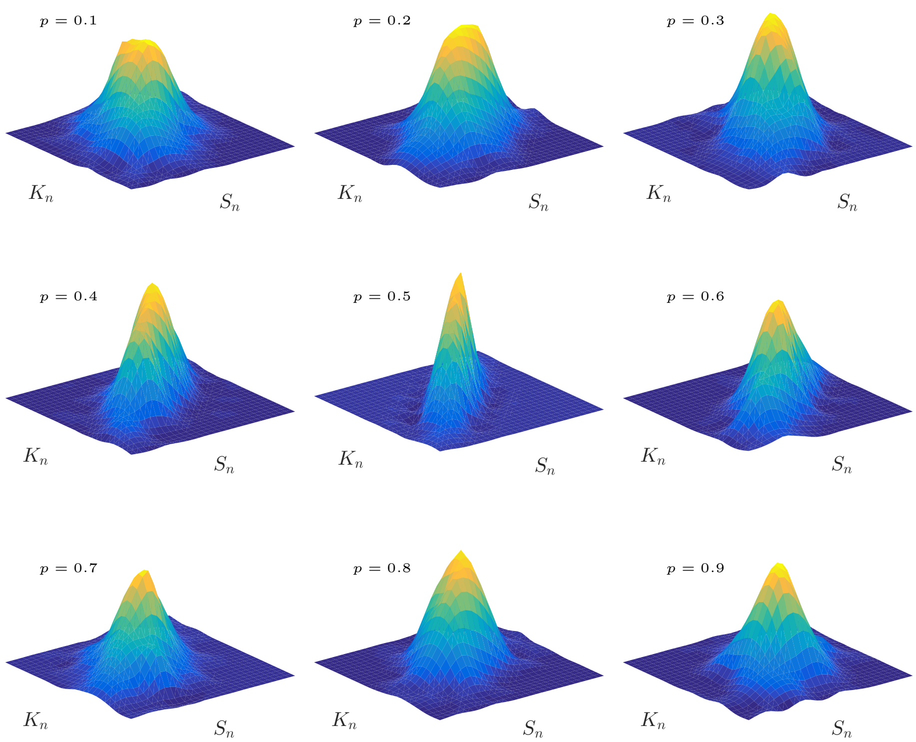

Fig. 3: *Joint distributions of* (S_n, K_n) *by Monte-Carlo simulations for* $n = 10^7$ *and varying* p*: the case* $p = 0.5$ *is seen to have stronger dependence than the others.*

These results are to be compared with the corresponding ones for random m-ary search trees [1], and the differences for correlation coefficients are summarized in Table 2. Furthermore, the joint distribution for

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	$\begin{cases} 3 \leq m \leq 26 :\to 0 \\ m \geq 27 : \text{periodic} \end{cases}$	

Tab. 2: *A comparison of the correlation coefficients for random tries and 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.*

m-ary search trees undergoes a phase change at $m = 26$: if the branching factor m satisfies $3 \leq m \leq 26$, then the space requirement is asymptotically independent from KPL and NPL, while for $m \geq 27$, their limiting joint distributions contain periodic fluctuations and are dependent; see [1] for more information.

Finally, similar results as those in this paper also hold for other digital families of trees, but for simplicity we focus on tries in this paper; see [7, 4] for more references.

2 Covariance and Correlation Coefficient

In this section, we sketch the main ideas leading to the proof of Theorem A on the asymptotics of the covariance and correlation coefficient of S_n and K_n. For the latter, we also need the variances of S_n and K_n which have been known for a long time; see Jacquet and Régnier [8], Kirschenhofer and Prodinger [10], Kirschenhofer et al. [11], Régnier and Jacquet [14] or the recent paper [4]. (See also Table 1 for a 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 [8]. See Flajolet et al. [3] for a survey on Mellin transform, and Jacquet and Szpankowski [9] 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 [4, 7].

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 $\mathbb{E}(K_n)$: $\tilde{f}_{1,0}(z) := e^{-z}\sum_{n\geq 0}\mathbb{E}(S_n)\frac{z^n}{n!}$ and $\tilde{f}_{0,1}(z) := e^{-z}\sum_{n\geq 0}\mathbb{E}(K_n)\frac{z^n}{n!}$. 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 [3],

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

for large $|z|$ in the half-plane $\Re(z) \geq \varepsilon > 0$, where h denotes the entropy of Bernoulli(p). Then, by Cauchy's integral representation and analytic de-Poissonization techniques [9], we obtain precise asymptotic approximations to $\mathbb{E}(S_n)$ and to $\mathbb{E}(K_n)$.

Similarly, for the variances $\mathbb{V}(S_n)$ and $\mathbb{V}(K_n)$, we introduce the Poisson generating functions of the second moments: $\tilde{f}_{2,0}(z) := e^{-z}\sum_{n\geq 0}\mathbb{E}(S_n^2)\frac{z^n}{n!}$ and $\tilde{f}_{0,2}(z) := e^{-z}\sum_{n\geq 0}\mathbb{E}(K_n^2)\frac{z^n}{n!}$, 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. It remains to derive the claimed estimate for the covariance. For that purpose, we then introduce the Poisson generating function $\tilde{f}_{1,1}(z) := e^{-z}\sum_{n\geq 0}\mathbb{E}(S_nK_n)\frac{z^n}{n!}$, which satisfies, again by (1),

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

To compute the covariance, it is beneficial to introduce now the *Poissonized covariance* (see (7) or [4] 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

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

and

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

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 $\mathscr{S} = \{z \in \mathbb{C} \,:\, \Re(z) \geq \epsilon, |\mathrm{Arg}(z)| \leq \theta_0\}$ 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 $\langle -2, 0\rangle$, and we have then the inverse Mellin integral representation

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

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

We then show that $\mathscr{M}[\tilde{h}_1(z); s]$ can be analytically continued to the vertical line $\Re(s) = -1$ and has no singularities there. This is the most complicated 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 $\mathscr{M}[\tilde{h}_1(z); s]$ becomes a Mellin convolution integral. In [4], a general procedure was given for the simplification of such integrals (see [4, p. 24 *et seq.*]). This simplification procedure and a direct application of the theory of admissible functions of analytic de-Poissonization now yield

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

$$\begin{aligned} g_k^{(2)} &= \frac{\Gamma(\chi_k)}{h}\Big(1 - \frac{\chi_k + 2}{2^{\chi_k+1}}\Big) - \frac{1}{h^2}\sum_{j\in\mathbb{Z}\setminus\{0\}} \Gamma(\chi_{k-j}+1)(\chi_j - 1)\Gamma(\chi_j) \\ &\quad - \frac{\Gamma(\chi_k+1)}{h^2}\Big(\gamma + 1 + \psi(\chi_k+1) - \frac{p\log^2 p + q\log^2 q}{2h}\Big) \\ &\quad + \frac{1}{h}\sum_{\ell\ge 2} \frac{(-1)^\ell (p^\ell + q^\ell)}{\ell!(1 - p^\ell - q^\ell)}\Gamma(\chi_k + \ell - 1)(2\ell^2 - 2\ell + 1 + \chi_k(2\ell - 1)), \end{aligned} \tag{9}$$

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 relevant and the second term (the sum over j) on the right-hand side of* (9) *has to be dropped. Also the first term here $\frac{\Gamma(\chi_k)}{h}\big(1 - \frac{\chi_k+2}{2^{\chi_k+1}}\big)$ 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 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 [4]. 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 [15] for details.

When $p = \frac{1}{2}$, an alternative expression to (9) (avoiding the convolution of two Fourier series) is

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

see the discussion of the size of tries in [4], where a similar alternative expression was given for $g_k^{(1)}$, which reads

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

Moreover, also in [4], the following expression for $g_k^{(3)}$ can be found

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

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 skipped here. The key tool of the proof is the multivariate version of the contraction method; see Neininger and Rüschendorf [13]. More precisely, we will use Theorem 3.1 in [13].

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

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

with the inverse

$$M^{-\frac{1}{2}} = \frac{1}{\sqrt{(ac - b^2)\big(a + c + 2\sqrt{ac - b^2}\big)}} \begin{pmatrix} c + \sqrt{ac - b^2} & -b \\ -b & a + \sqrt{ac - b^2} \end{pmatrix}. \tag{10}$$

Now we sketch the proof of Theorem B, Part (i).

Proof of Theorem B, Part (i). First note that

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

where the notation is as in Section 1. The contraction method was specially developed for obtaining limiting distribution results for such recurrences; see [13].

We need some notation. First, define

$$\widehat{\Sigma}_n := \begin{pmatrix} \mathbb{V}(S_n) & \mathrm{Cov}(S_n, K_n) \\ \mathrm{Cov}(S_n, K_n) & \mathbb{V}(K_n) \end{pmatrix}. \tag{11}$$

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 [13], it suffices to show that the following conditions hold

$$b_n^{(i)} \xrightarrow{L_3} 0, \qquad M_n^{(i)} \xrightarrow{L_3} M_i, \tag{12}$$

$$\mathbb{E}\big(\|M_1\|_{\text{op}}^3 + \|M_2\|_{\text{op}}^3\big) < 1, \qquad \mathbb{E}\big(\|M_n^{(i)}\|_{\text{op}}^3 \chi_{\{B_n^{(i)} \le j\} \cup \{B_n^{(i)} = n\}}\big) \longrightarrow 0 \tag{13}$$

for $i = 1, 2$ and $j \in \mathbb{N}$, where $\xrightarrow{L_3}$ denotes convergence in the L_3-norm, $\|\cdot\|_{\text{op}}$ 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 [13] 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

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

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 (12) because the second condition of (13) follows along similar lines and the first condition of (13) follows from (12) in view of

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

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

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

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 (10), we see that

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

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 [5] and a similar method of proof also applies to K_n), we have

$$\frac{1-\mu(n)+\mu(B_n)+\mu(n-B_n)}{\sqrt{n}} \xrightarrow{L_3} 0 \qquad \text{and} \qquad \frac{n-\nu(n)+\nu(B_n)+\nu(n-B_n)}{\sqrt{n\log n}} \xrightarrow{L_3} 0.$$

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 the claimed result 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 of the matrix square-root and its inverse from above. □

From this lemma and Proposition B our claimed result now follows.

Acknowledgments

We thank the referees for helpful comments; in particular, the detailed comments by one referee on the proof of Theorem B led to considerable improvements in Section 3.

References

[1] H.-H. Chern, M. Fuchs, H.-K. Hwang and R. Neininger. Dependence and phase changes in random m-ary search trees, *Random Struct. Algor.*, to appear.

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

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

[4] 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.

[5] 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.*, **21:1**, 26 pages.

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

[7] 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.

[8] 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.

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

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

[11] 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.

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

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

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

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

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