Protected node profile of Tries

Mehri Javanian

Protected node profile of Tries

Mehri Javanian[1*]

[1] *Department of Statistics, University of Zanjan, Iran*

In a rooted tree, protected nodes are neither leaves nor parents of any leaves. They have some practical motivations, *e.g.*, in organizational schemes, security models and social-network models. Protected node profile measures the number of protected nodes with the same distance from the root in rooted trees. For no rooted tree, protected node profile has been investigated so far. Here, we present the asymptotic expectations, variances, covariance and limiting bivariate distribution of protected node profile and non-protected internal node profile in random tries, an important data structure on words in computer science. Also we investigate the fraction of these expectations asymptotically. These results are derived by the methods of analytic combinatorics such as generating functions, Mellin transform, Poissonization and depoissonization, saddle point method and singularity analysis.

Keywords: Tries, Protected nodes, Tree profiles, Poissonization, Mellin transform, Recurrences, Generating functions, Singularity analysis, Saddle point method

1 Introduction

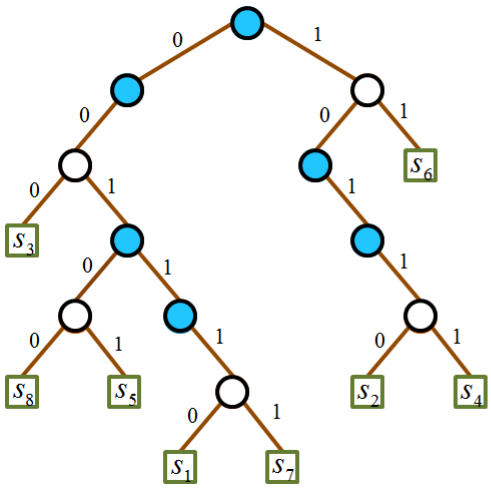
Tries (invented by de la Briandais) are fundamental tree data structures for re*trie*val of information. The information stored in a trie is a set of strings (see Knuth (1998) for more details). For simplicity, we consider 0-1 strings to store in a trie. In a trie, the subject of this paper, strings are stored in leaves. More precisely, a trie is built on n infinite 0-1 strings as follows: if $n = 1$ then the only string is stored in the root as an external node; if $n > 1$, then the root is an internal node (empty node) and the strings with the first bit "0" ("1") are directed to the left (right) subtree; finally, the subtrees are constructed recursively by the same rules, but by removing the first bit of all strings (*cf.* Figure 1).

A random trie with n external nodes is a trie built over n infinite 0-1 strings (a trie of size n) generated by memoryless source, that is, we assume each string is a Bernoulli i.i.d. sequence with success probability $0 < p < 1$ (the probability of occurring a "1"); we also use $q := 1 - p \leq p$. Random tries have been extensively studied; for more background, see Mahmoud (1992) or the survey Park et al. (2009), and the references therein, for a thorough analysis of the profile (number of nodes at a given level) of tries.

By protected nodes, we mean the nodes with a distance of at least two to all the leaves. *E.g.*, Figure 1 shows the protected nodes in blue color. Protected nodes were introduced by Cheon and Shapiro (2008) as a guide in various organizational schemes. For instance, if leaves represent customers it may be worthwhile for many of the points in the tree to be non-protected. However if the leaves represent lobbyists or computer hackers it may be a very good thing to have many points protected. In a security model with trie

*Email: javanian@znu.ac.ir

ISSN subm. to DMTCS



$$I_{8,0} = 1,\quad X_{8,0} = 1,\quad Y_{8,0} = 0$$
$$I_{8,1} = 2,\quad X_{8,1} = 1,\quad Y_{8,1} = 1$$
$$I_{8,2} = 2,\quad X_{8,2} = 1,\quad Y_{8,2} = 1$$
$$I_{8,3} = 2,\quad X_{8,3} = 2,\quad Y_{8,3} = 0$$
$$I_{8,4} = 3,\quad X_{8,4} = 1,\quad Y_{8,4} = 2$$
$$I_{8,5} = 1,\quad X_{8,5} = 0,\quad Y_{8,5} = 1$$

Fig. 1: A trie built on eight strings $s_1, \ldots, s_8$ (*i.e.*, $s_1 = 001110\ldots$, $s_2 = 10110\ldots$, $s_3 = 000\ldots$, $s_4 = 10111\ldots$, *etc.*) with internal (circles), leaf (squares), protected (blue circles), and non-protected (white circles) nodes, and its profiles.

structure, a protected node may be taken to represent an entity that has at least two buffers between itself and a vulnerable point. Protected nodes have been investigated for many different random trees by many authors; see for instance, Du and Prodinger (2012), Devroye and Janson (2014), Fuchs et al. (2016) and the papers cited there.

For random tries, the mean and variance of the number of protected nodes have been obtained by Gaither et al. (2012) and Gaither and Ward (2013) where the applications of this parameter such as security models with trie structures and social networks have been discussed. Moreover, Gaither and Ward (2013) announced a central limit theorem, which was conjectured in their paper. This conjecture has been confirmed by Fuchs et al. (2016); with proving (univariate and bivariate) central limit theorems for the number of protected nodes. Also, Fuchs et al. (2016) have shown the all previous results can be derived by approaches of Hwang et al. (2012), Fuchs et al. (2014) and Fuchs and Lee (2014).

In the present paper, we are concerned with the protected node profile defined as the number of protected nodes with the same distance from the root in random tries. Throughout the paper, we write $I_{n,k}$, $X_{n,k}$ and $Y_{n,k}$ for, respectively, the number of internal nodes, the number of protected nodes and the number of non-protected internal nodes at level k in a trie of size n. Namely, $I_{n,k} = X_{n,k} + Y_{n,k}$ (*cf.* Figure 1).

The paper is organized as follows. In the next section, we show that the probability generating functions of $X_{n,k}$ and $Y_{n,k}$, satisfy a recurrence of the form

$$z_{n,k}(u) = \sum_{l=0}^{n} \binom{n}{l} p^l q^{n-l} z_{l,k-1}(u) z_{n-l,k-1}(u), \qquad (n \geq 0, k \geq 1).$$

Therefore the expectations, the variances and covariance of $X_{n,k}$ and $Y_{n,k}$, satisfy a recurrence of the

form

$$t_{n,k} = \sum_{l=0}^{n} \binom{n}{l} p^l q^{n-l} \Big(t_{l,k-1} + t_{n-l,k-1} \Big), \qquad (n \geq 0, k \geq 1). \tag{1}$$

In Sections 3 and 4, in order to derive the asymptotic approximations to the average profiles, variances and covariance, we use a standard approach: first we consider the Poisson generating function $f_k(z) := e^{-z} \sum_n t_{n,k} z^n/n!$, which, by (1) satisfies the functional equation

$$f_k(z) = f_{k-1}(pz) + f_{k-1}(qz).$$

This equation can be solved explicitly by a simple iteration argument and asymptotically by using the Mellin transform (see Flajolet et al. (1995) and Szpankowski (2001)). The final step is to invert from the asymptotics of the Poisson generating function $f_k(z)$ to recover the asymptotics of $t_{n,k}$. This last step is guided by the Poisson heuristic, which roughly states that

if a sequence $\{a_n\}_n$ *is "smooth enough", then* $a_n \sim e^{-n} \sum_{j \geq 0} a_n n^j/j!$,

where $a_n \sim b_n$ if $\lim_{n\to\infty} a_n/b_n = 1$. This Poisson heuristic is known as analytic de-Poissonization, when justified by complex analysis and the saddle-point method. Our results show that for ($\varepsilon > 0$)

$$\frac{1}{\log(1/q)} + \varepsilon \leq \frac{2}{\log(1/p) + \log(1/q)} + \varepsilon \leq k \leq \frac{p^2 + q^2}{p^2 \log(1/p) + q^2 \log(1/q)} - \varepsilon, \tag{2}$$

and

$$\frac{1}{\log(1/q)} + \varepsilon \leq k \leq \frac{p^2 + q^2}{p^2 \log(1/p) + q^2 \log(1/q)} - \varepsilon,$$

respectively, oscillating factors emerge in $\mathbb{E}(X_{n,k})$ and $\mathbb{E}(Y_{n,k})$. Such a behavior is a consequence of an infinite number of saddle-points appearing in the integrand of the associated Mellin integral transform. This was first observed by Nicodème (2005). Then we investigate the ratio $\mathbb{E}(X_{n,k})/\mathbb{E}(Y_{n,k})$ for the range of k in (2). We do not consider other ranges of k, because for $k \leq 1/\log(1/q)$, each level is almost full of internal nodes, and for $k \geq (p^2 + q^2)/(p^2 \log(1/p) + q^2 \log(1/q))$, $\mathbb{E}(X_{n,k})$ and $\mathbb{E}(Y_{n,k})$ tend to zero (see Park et al. (2009)). Also we prove that the variances of both profiles are asymptotically of the same order as their expected values. We then show, in Section 5, that $X_{n,k}$ and $Y_{n,k}$, after a proper normalization, have a bivariate normal limiting joint distribution for the range of k in (2), if and only if the variances tend to infinity.

Here, we focus mostly on the protected node and non-protected node profiles of asymmetric tries (when $p \neq q$) since the symmetric tries (when $p = q = 1/2$) are comparatively easier. In the last section, we briefly summarize our main results for symmetric tries.

2 Preliminaries

In a random trie of size n, the number of protected nodes at level $k \geq 1$, $X_{n,k}$ can be computed recursively by computing the number for the two subtrees at level $k-1$. For $k = 0$, the root is protected, if and only

if neither the left nor the right subtree contains only one string. This leads to the following distributional recurrence for $X_{n,k}$:

$$X_{n,k} \stackrel{d}{=} \begin{cases} X_{B_n,k-1} + X^*_{n-B_n,k-1}, & k \geq 1; \\ 1 - \mathbb{I}_{\{1,n-1\}}(B_n), & k = 0, \end{cases} \qquad (n \geq 2),$$

where $\mathbb{I}_A(\cdot)$ is the indicator function of A, $X_{n,k} \stackrel{d}{=} X^*_{n,k}$, $B_n \stackrel{d}{=}$ Binomial(n,p) and $X_{n,k}$, $X^*_{n,k}$, B_n are independent. Also, for $k \geq 0$, $X_{0,k} = X_{1,k} = 0$.

Similarly, we have $Y_{0,k} = Y_{1,k} = 0$ for $k \geq 0$, and

$$Y_{n,k} \stackrel{d}{=} \begin{cases} Y_{B_n,k-1} + Y^*_{n-B_n,k-1}, & k \geq 1; \\ \mathbb{I}_{\{1,n-1\}}(B_n), & k = 0, \end{cases} \qquad (n \geq 2),$$

where $Y_{n,k} \stackrel{d}{=} Y^*_{n,k}$ and $Y_{n,k}$, $Y^*_{n,k}$, B_n are independent.

Let $P^{[X]}_{n,k}(u) := \mathbb{E}[u^{X_{n,k}}]$ and $P^{[Y]}_{n,k}(u) := \mathbb{E}[u^{Y_{n,k}}]$. Then $P^{[X]}_{n,k}(u)$ and $P^{[Y]}_{n,k}(u)$ are both solutions to the following recurrence relation with respect to $z_{n,k}(u)$:

$$z_{n,k}(u) = \sum_{l=0}^{n} \binom{n}{l} p^l q^{n-l} z_{l,k-1}(u) z_{n-l,k-1}(u), \qquad (n \geq 0, k \geq 1), \tag{3}$$

with the initial and boundary conditions

$$P^{[X]}_{n,0}(u) = \begin{cases} u - n(pq^{n-1} + p^{n-1}q)(u-1), & n \geq 3; \\ u - 2pq(u-1), & n = 2, \\ 1, & n = 0, 1, \end{cases}$$

$$P^{[Y]}_{n,0}(u) = \begin{cases} 1 + n(pq^{n-1} + p^{n-1}q)(u-1), & n \geq 3; \\ 1 + 2pq(u-1), & n = 2, \\ 1, & n = 0, 1. \end{cases}$$

Throughout the paper, we use the following notations. For a complex number s we define the function

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

For a real number α, the equation

$$\alpha = \frac{p^{-\rho} + q^{-\rho}}{p^{-\rho}\log(1/p) + q^{-\rho}\log(1/q)},$$

that $(\log \frac{1}{q})^{-1} < \alpha < (\log \frac{1}{p})^{-1}$, has the solution

$$\rho = \rho(\alpha) = \frac{1}{\log(p/q)} \log \frac{1 - \alpha\log(1/p)}{\alpha\log(1/q) - 1}. \tag{4}$$

We also define the functions

$$\alpha_1 = \frac{1}{\log(1/q)}, \qquad \alpha_0 = \frac{2}{\log(1/p) + \log(1/q)},$$

$$\alpha_2 = \frac{p^2 + q^2}{p^2\log(1/p) + q^2\log(1/q)}, \qquad \beta(\rho) = \frac{p^{-\rho}q^{-\rho}\log(p/q)^2}{(p^{-\rho} + q^{-\rho})^2}.$$

3 Expectations of $X_{n,k}$, $Y_{n,k}$ and Their Ratio

Asymptotic approximations to the expectations of $X_{n,k}$, $Y_{n,k}$ are derived in this section. Also, we give some result about the value of $\mathbb{E}(X_{n,k})/\mathbb{E}(Y_{n,k})$.

Let $\mu_{n,k}^{[X]} := \mathbb{E}(X_{n,k})$ and $\mu_{n,k}^{[Y]} := \mathbb{E}(Y_{n,k})$. Then from (3), $\mu_{n,k}^{[X]}$ and $\mu_{n,k}^{[Y]}$ are both solutions to the following recurrence with respect to $\mu_{n,k}$:

$$\mu_{n,k} = \sum_{j=0}^{n} \binom{n}{j} p^j q^{n-j} \big(\mu_{j,k-1} + \mu_{n-j,k-1}\big), \qquad (n \geq 0, k \geq 1),$$

with the initial and boundary conditions

$$\mu_{n,0}^{[X]} = \begin{cases} 1 - n(pq^{n-1} + p^{n-1}q), & n \geq 3; \\ 1 - 2pq, & n = 2, \\ 0, & n = 0, 1, \end{cases} \qquad \mu_{n,0}^{[Y]} = \begin{cases} n(pq^{n-1} + p^{n-1}q), & n \geq 3; \\ 2pq, & n = 2, \\ 0, & n = 0, 1. \end{cases}$$

It follows that the poisson transforms

$$M_k^{[X]}(x) := \sum_{n\geq 0} \mu_{n,k}^{[X]} \frac{x^n}{n!} e^{-x} \quad \text{and} \quad M_k^{[Y]}(x) := \sum_{n\geq 0} \mu_{n,k}^{[Y]} \frac{x^n}{n!} e^{-x},$$

satisfy

$$M_k^{[X]}(x) = \sum_{j=0}^{k} \binom{k}{j} M_0^{[X]}(p^j q^{k-j} x) = 2^k - \sum_{j=0}^{k} \binom{k}{j} \hat{M}_0^{[X]}(p^j q^{k-j} x) := 2^k - \hat{M}_k^{[X]}(x), \tag{5}$$

$$M_k^{[Y]}(x) = \sum_{j=0}^{k} \binom{k}{j} M_0^{[Y]}(p^j q^{k-j} x), \tag{6}$$

for $k \geq 1$ with initial conditions

$$\begin{aligned} M_0^{[X]}(x) &= 1 - e^{-x} - pxe^{-px} - qxe^{-qx} + pqx^2 e^{-x} := 1 - \hat{M}_0^{[X]}(x), \\ M_0^{[Y]}(x) &= pxe^{-px} + qxe^{-qx} - xe^{-x} - pqx^2 e^{-x}. \end{aligned}$$

Let $M_k^{*[X]}(s)$ and $M_k^{*[Y]}(s)$ denote the Mellin transforms

$$M_k^{*[X]}(s) = \int_0^\infty M_k^{[X]}(x) x^{s-1} dx \quad \text{and} \quad M_k^{*[Y]}(s) = \int_0^\infty M_k^{[Y]}(x) x^{s-1} \mathrm{d}x,$$

that $M_k^{*[X]}(s)$ exists for $s \in \mathbb{C}$ with $-2 < \Re(s) < 0$; and $M_k^{*[Y]}(s)$ exists for $s \in \mathbb{C}$ with $\Re(s) > -2$. Then (5) and (6) rewrite to

$$\hat{M}_k^{*[X]}(s) = (p^{-s} + q^{-s})^k \hat{M}_0^{*[X]}(s), \qquad M_k^{*[Y]}(s) = (p^{-s} + q^{-s})^k M_0^{*[Y]}(s),$$

with initial conditions

$$\hat{M}_0^{*[X]}(s) = \Gamma(s)\big(1 + s(p^{-s} + q^{-s}) - s(s+1)pq\big),$$
$$M_0^{*[Y]}(0) = \Gamma(s+1)\big(p^{-s} + q^{-s} - 1 - (s+1)pq\big).$$

Hence, by the inverse Mellin transform (Flajolet et al. (1995))

$$M_k^{[X]}(x) = 2^k - \frac{1}{2\pi i}\int_{\rho-i\infty}^{\rho+i\infty} \Gamma(s)\hat{g}^{[X]}(s)(p^{-s}+q^{-s})^k x^{-s}\mathrm{d}s, \qquad -2 < \rho < 0, \tag{7}$$

$$M_k^{[Y]}(x) = \frac{1}{2\pi i}\int_{\rho-i\infty}^{\rho+i\infty} \Gamma(s+1)g^{[Y]}(s)(p^{-s}+q^{-s})^k x^{-s}\mathrm{d}s, \qquad \rho > -2, \tag{8}$$

where $\rho := \Re(s)$ and

$$\hat{g}^{[X]}(s) = 1 + s(p^{-s}+q^{-s}) - s(s+1)pq := -g^{[X]}(s),$$
$$g^{[Y]}(s) = p^{-s} + q^{-s} - 1 - (s+1)pq.$$

We are mainly interested in the behaviour of $M_k^{[X]}(x)$ and $M_k^{[Y]}(x)$ for $x = n$, since by analytic depoissonization we expect that $\mathbb{E}(X_{n,k}) \sim M_k^{[X]}(n)$ and $\mathbb{E}(Y_{n,k}) \sim M_k^{[Y]}(n)$.

Here, we evaluate the integrals (7) and (8) via the saddle point method (see Szpankowski (2001)). Hence, it is natural to choose $\rho = \rho_{n,k}$ as the saddle point of the function

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

that is the solution of the equation $\frac{\partial}{\partial s}(k\log T(s) - s\log n) = 0$. Equivalently we must find ρ from

$$\frac{k}{\log n} = \frac{p^{-\rho} + q^{-\rho}}{p^{-\rho}\log(1/p) + q^{-\rho}\log(1/q)}, \tag{9}$$

that is, the only real-valued saddle point $\rho = \rho_{n,k} = \rho(\frac{k}{\log n})$ (see (4)).

The integrands in (7) and (8), also has infinitely many complex-valued saddle points of the form $s_j := \rho + 2\pi i j/(\log p/q)$ $(j = \pm1, \pm2, \ldots)$. This is due to the fact

$$T(\rho + it) = p^{-\rho-it}\left(1 + \big(\frac{q}{p}\big)^{-\rho-it}\right) = p^{-\rho}\cdot e^{-it\log p}\left(1 + \big(\frac{q}{p}\big)^{-\rho}\cdot e^{it\log\frac{p}{q}}\right).$$

Now by putting $t = 2\pi j/(\log p/q)$, we have

$$\begin{aligned} T\big(\rho + 2\pi i j/(\log p/q)\big) &= p^{-\rho}\cdot e^{-2\pi i j(\log p)/(\log p/q)}\left(1 + \big(\frac{q}{p}\big)^{-\rho}\cdot e^{2\pi i j}\right) \\ &= e^{-2\pi i j(\log p)/(\log p/q)}T(\rho). \end{aligned}$$

Consequently the behaviour of $T(s)^k x^{-s}$ around $s = s_j$ is almost the same as that of $T(s)^k x^{-s}$ around $s = \rho$. This phenomenon gives a periodic leading factor in the asymptotics of $M_k^{[X]}(n)$ and $M_k^{[Y]}(n)$; and also of $\mu_{n,k}^{[X]} = \mathbb{E}(X_{n,k})$ and $\mu_{n,k}^{[Y]} = \mathbb{E}(Y_{n,k})$.

Theorem 3.1 *Consider $\rho_{n,k} = \rho(k/\log n)$ and $t_j = 2\pi j/(\log p/q)$. For some $\varepsilon > 0$,*

1. *If $\alpha_1 + \varepsilon \le \frac{k}{\log n} \le \alpha_0 - \varepsilon$ then*

$$\mu_{n,k}^{[X]} = 2^k - \hat{G}^{[X]}\left(\rho_{n,k}, \log_{p/q} p^k n\right) \frac{(p^{-\rho_{n,k}} + q^{-\rho_{n,k}})^k n^{-\rho_{n,k}}}{\sqrt{2\pi\beta(\rho_{n,k})k}} \left(1 + \mathcal{O}(k^{-1/2})\right),$$

where $\hat{G}^{[X]}(\rho, x) = \sum_{j\in\mathbb{Z}} \Gamma(\rho + it_j)\hat{g}^{[X]}(\rho + it_j)e^{-2\pi ijx}$ is a non-zero 1-periodic function.

2. *If $\alpha_0 + \varepsilon \le \frac{k}{\log n} \le \alpha_2 - \varepsilon$ then*

$$\mu_{n,k}^{[X]} = G^{[X]}\left(\rho_{n,k}, \log_{p/q} p^k n\right) \frac{(p^{-\rho_{n,k}} + q^{-\rho_{n,k}})^k n^{-\rho_{n,k}}}{\sqrt{2\pi\beta(\rho_{n,k})k}} \left(1 + \mathcal{O}(k^{-1/2})\right),$$

where $G^{[X]}(\rho, x) = \sum_{j\in\mathbb{Z}} \Gamma(\rho + it_j)g^{[X]}(\rho + it_j)e^{-2\pi ijx}$ is a non-zero 1-periodic function.

3. *If $\alpha_1 + \varepsilon \le \frac{k}{\log n} \le \alpha_2 - \varepsilon$ then*

$$\mu_{n,k}^{[Y]} = G^{[Y]}\left(\rho_{n,k}, \log_{p/q} p^k n\right) \frac{(p^{-\rho_{n,k}} + q^{-\rho_{n,k}})^k n^{-\rho_{n,k}}}{\sqrt{2\pi\beta(\rho_{n,k})k}} \left(1 + \mathcal{O}(k^{-1/2})\right),$$

where $G^{[Y]}(\rho, x) = \sum_{j\in\mathbb{Z}} \Gamma(\rho + it_j + 1)g^{[Y]}(\rho + it_j)e^{-2\pi ijx}$ is a non-zero 1-periodic function.

Proof: By evaluating the integrals (7) and (8) via the saddle point method, the proof is quite identical to that of Theorem 2 in Park et al. (2009) (Lemma 7.5 in Drmota (2009)) with the new functions $\hat{g}^{[X]}(s)$, $g^{[X]}(s)$ and $g^{[Y]}(s)$. □

Remark 1 *For the sake of the following global picture of a random trie, we only consider the three ranges of k in Theorem 3.1:*

- *When $1 \le \frac{k}{\log n} \le \alpha_1 - \varepsilon$, each level is almost full of internal nodes, ($I_{n,k} \approx 2^k$, $X_{n,k} \approx 2^k$, $Y_{n,k} \approx 0$); in particular, the variances of profiles, $\mathbb{V}(I_{n,k})$, $\mathbb{V}(X_{n,k})$ and $\mathbb{V}(Y_{n,k})$ tend to zero.*

- *When $\alpha_1 + \varepsilon \le \frac{k}{\log n} \le \alpha_0 - \varepsilon$, $\rho_{n,k} = \rho(k/\log n) > 0$; and when $\alpha_0 + \varepsilon \le \frac{k}{\log n} \le \alpha_2 - \varepsilon$, $-2 < \rho_{n,k} = \rho(k/\log n) < 0$. For the second case, $\mathbb{V}(X_{n,k}) \to \infty$, $\mathbb{V}(Y_{n,k}) \to \infty$, and we prove the asymptotic bivariate normality of $X_{n,k}$ and $Y_{n,k}$.*

- *When $\frac{k}{\log n} \ge \alpha_2 + \varepsilon$, then $\mathbb{E}(I_{n,k})$ and $\mathbb{E}(X_{n,k})$ tend to zero.*

Tab. 1: Comparisons of magnitudes for $x \in (0,1)$ (Since $\alpha_0 \leq \alpha \leq \alpha_2$ then $-2 < \rho(\alpha) < 0$).

Functions of x	p	$\rho(\alpha)$ -0.1	-1	-1.9	Oscillation
$G^{[Y]}(\rho(\alpha), x)$	0.55	0.68	0.44	2.50	Almost flat
	0.75	0.72	0.37	1.91	Non-flat
	0.95	0.73	0.15	0.49	Non-flat
$G^{[X]}(\rho(\alpha), x)$	0.55	8.90	0.56	2.50	Almost flat
	0.75	8.89	0.63	3.10	Non-flat
	0.95	8.85	0.85	4.50	Non-flat
$\frac{G^{[X]}(\rho(\alpha),x)}{G^{[Y]}(\rho(\alpha),x)}$	0.55	13.09	1.27	1.00	Almost flat
	0.75	12.35	1.66	1.62	Non-flat
	0.95	12.12	5.00	9.10	Non-flat

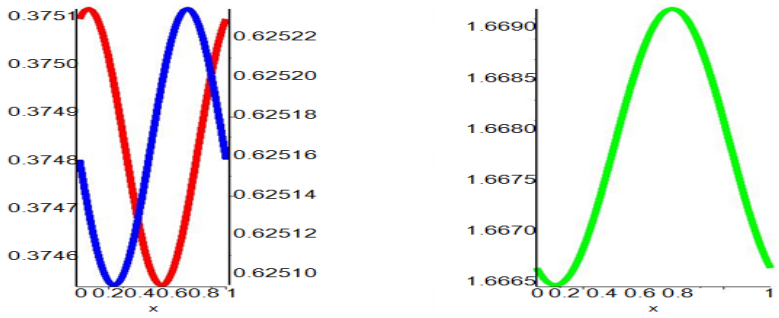


Fig. 2: The fluctuating part of the functions $G^{[Y]}(-1,x)$ (red curve), $G^{[X]}(-1,x)$ (blue curve) and $G^{[X]}(-1,x)/G^{[Y]}(-1,x)$ (green curve), for $x \in (0,1)$ and $p = 0.75$.

Theorem 3.2 *Let* $\alpha_{n,k} := \frac{k}{\log n}$. *When* $p \to \frac{1}{2}^+$, *then* $\rho_0(\alpha_{n,k}) := \lim_{p\to\frac{1}{2}^+} \rho(\alpha_{n,k}) = \frac{\alpha_{n,k}}{1-\alpha_{n,k}\log 2}$ *and*

$$\frac{\mu_{n,k}^{[X]}}{\mu_{n,k}^{[Y]}} \longrightarrow \begin{cases} \infty, & \text{if } \alpha_{n,k} \to {\alpha_0}^+; \\ \frac{\rho_0(\alpha_{n,k})\left(1+\rho_0(\alpha_{n,k})\right)-8\rho_0(\alpha_{n,k})2^{\rho_0(\alpha_{n,k})}-4}{8\rho_0(\alpha_{n,k})2^{\rho_0(\alpha_{n,k})}-4\rho_0(\alpha_{n,k})-\rho_0(\alpha_{n,k})\left(1+\rho_0(\alpha_{n,k})\right)}, & \text{if } \alpha_0 < \alpha_{n,k} < \alpha_2; \\ 1, & \text{if } \alpha_{n,k} \to {\alpha_2}^-. \end{cases}$$

Proof: By (4), it is easy to see that $\rho_0(\alpha) = \frac{\alpha}{1-\alpha\log 2}$. Then from Theorem 3.1, as $p \to \frac{1}{2}^+$,

$$\frac{\mu_{n,k}^{[X]}}{\mu_{n,k}^{[Y]}} = \frac{G^{[X]}\left(\rho_{n,k}, \log_{p/q} p^k n\right)}{G^{[Y]}\left(\rho_{n,k}, \log_{p/q} p^k n\right)} \longrightarrow \frac{\Gamma\left(\rho_0(\alpha_{n,k})\right) g^{[X]}(\rho_0(\alpha_{n,k}))}{\Gamma\left(\rho_0(\alpha_{n,k})+1\right) g^{[Y]}(\rho_0(\alpha_{n,k}))} = \frac{g^{[X]}\left(\rho_0(\alpha_{n,k})\right)}{\rho_0(\alpha_{n,k}) g^{[Y]}\left(\rho_0(\alpha_{n,k})\right)}.$$

Substituting $g^{[X]}\left(\rho_0(\alpha_{n,k})\right)$ and $g^{[Y]}\left(\rho_0(\alpha_{n,k})\right)$, we obtain the result. □

Table 1 and Figure 2 give several examples that confirm the claims in Theorem 3.2.

4 Variances and Covariance

Asymptotic approximations to the variances of $X_{n,k}$, $Y_{n,k}$ and their covariance, $\sigma_{n,k}^{[X]^2} := \mathbb{V}(X_{n,k})$, $\sigma_{n,k}^{[Y]^2} := \mathbb{V}(Y_{n,k})$ and $\gamma_{n,k} := \mathrm{Cov}(X_{n,k}, Y_{n,k})$ respectively, are derived in this section.

Let $N_k^{[X]}(x) := \sum_{n\geq 0} \mathbb{E}(X_{n,k}^2)\frac{x^n}{n!}e^{-x}$ and $N_k^{[Y]}(x) := \sum_{n\geq 0} \mathbb{E}(Y_{n,k}^2)\frac{x^n}{n!}e^{-x}$. Now define the Poisson variances, $V_k^{[X]}(x) := N_k^{[X]}(x) - M_k^{[X]^2}(x)$ and $V_k^{[Y]}(x) := N_k^{[Y]}(x) - M_k^{[Y]^2}(x)$. Then from (3) and similar to (5) and (6), for $k \geq 1$, it yields

$$V_k^{[X]}(x) = \sum_{j=0}^{k} \binom{k}{j} V_0^{[X]}(x)(p^j q^{k-j} x), \qquad V_k^{[Y]}(x) = \sum_{j=0}^{k} \binom{k}{j} V_0^{[Y]}(x)(p^j q^{k-j} x), \tag{10}$$

with initial conditions

$$\begin{aligned}
V_0^{[X]}(x) &= N_0^{[X]}(x) - M_0^{[X]^2}(x) = M_0^{[X]}(x) - M_0^{[X]^2}(x),\\
&= e^{-x} - e^{-2x} + pxe^{-px} + qxe^{-qx} - 2pxe^{-x(1+p)} - 2qxe^{-x(1+q)} + 2pqx^2e^{-2x}\\
&\quad -3pqx^2e^{-x} - p^2x^2e^{-2px} - q^2x^2e^{-2qx} + 2p^2qx^3e^{-x(1+p)} + 2pq^2x^3e^{-x(1+q)} - p^2q^2x^4e^{-2x},\\
V_0^{[Y]}(x) &= N_0^{[Y]}(x) - M_0^{[Y]^2}(x) = M_0^{[Y]}(x) - M_0^{[Y]^2}(x),\\
&= pxe^{-px} + qxe^{-qx} + 2px^2e^{-x(1+p)} + 2qx^2e^{-x(1+q)} + 2p^2qx^3e^{-x(1+p)} + 2pq^2x^3e^{-x(1+q)}\\
&\quad -p^2x^2e^{-2px} - q^2x^2e^{-2qx} - 2pqx^3e^{-2x} - p^2q^2x^4e^{-2x} - 3pqx^2e^{-x} - xe^{-x} - x^2e^{-2x}.
\end{aligned}$$

Consider $\bar{I}_{n,k} := 2^k - I_{n,k}$. By Section 6.2 in Park et al. (2009), for $k \geq 1$,

$$V_k^{[I]}(x) = \sum_{j=0}^{k} \binom{k}{j} V_0^{[I]}(x)(p^j q^{k-j} x), \quad \text{and} \quad V_0^{[I]}(x) = (x+1)e^{-x}(1-(x+1)e^{-x}),$$

where $V_k^{[I]}(x) := \sum_{n\geq 0} \mathbb{E}(\bar{I}_{n,k}^2)\frac{x^n}{n!}e^{-x} - \big(\sum_{n\geq 0} \mathbb{E}(\bar{I}_{n,k})\frac{x^n}{n!}e^{-x}\big)^2$. Hence, by (10) and for $k \geq 1$,

$$\begin{aligned}
2C_k(x) &= \sum_{j=0}^{k} \binom{k}{j} \Big(V_0^{[I]}(p^j q^{k-j} x) - V_0^{[X]}(p^j q^{k-j} x) - V_0^{[Y]}(p^j q^{k-j} x)\Big),\\
&= 2\sum_{j=0}^{k} \binom{k}{j} C_0(p^j q^{k-j} x), \qquad (C_0(x) := -M_0^{[X]}(x) M_0^{[Y]}(x)),
\end{aligned} \tag{11}$$

with $C_k(x) := \big(V_k^{[I]}(x) - V_k^{[X]}(x) - V_k^{[Y]}(x)\big)/2$, *i.e.* the Poisson covariance with initial condition

$$C_0(x) = -\big(pxe^{-px} + qxe^{-qx} - xe^{-x} - pqx^2e^{-x}\big)\big(1 - e^{-x} - pxe^{-px} - qxe^{-qx} + pqx^2e^{-x}\big).$$

Thus, for $\rho > -2$, we have

$$V_k^{[X]}(x) = \frac{1}{2\pi i}\int_{\rho-i\infty}^{\rho+i\infty} \Gamma(s+1) g_V^{[X]}(s)(p^{-s}+q^{-s})^k x^{-s}\mathrm{d}s, \tag{12}$$

$$V_k^{[Y]}(x) = \frac{1}{2\pi i}\int_{\rho-i\infty}^{\rho+i\infty} \Gamma(s) g_V^{[Y]}(s)(p^{-s}+q^{-s})^k x^{-s}\mathrm{d}s, \tag{13}$$

$$C_k(x) = \frac{1}{2\pi i}\int_{\rho-i\infty}^{\rho+i\infty} \Gamma(s) g_C(s)(p^{-s}+q^{-s})^k x^{-s}\mathrm{d}s, \tag{14}$$

where

$$\begin{aligned}
g_V^{[X]}(s) &= 1 - 2^{-s} + s\left(p^{-s}+q^{-s} - 2p(p+1)^{-s-1} - 2q(q+1)^{-s-1}\right)\\
&\quad -s(s+1)\left(2^{-s-2}(p^{-s}+q^{-s}) + 3pq - pq2^{-s-1}\right)\\
&\quad +2pqs(s+1)(s+2)\left(p(p+1)^{-s-3} + q(q+1)^{-s-3}\right)\\
&\quad -s(s+1)(s+2)(s+3)p^2q^2 2^{-s-4},\\
g_V^{[Y]}(s) &= p^{-s}+q^{-s} - 1 + (s+1)\left(2p(p+1)^{-s-2} + 2q(q+1)^{-s-2}\right)\\
&\quad -(s+1)\left(2^{-s-2}(p^{-s}+q^{-s}+1) + 3pq\right) - (s+1)(s+2)(s+3)p^2q^2 2^{-s-4}\\
&\quad +2pq(s+1)(s+2)\left(p(p+1)^{-s-3} + q(q+1)^{-s-3} - 2^{-s-3}\right),\\
g_C(s) &= s\left(p(p+1)^{-s-1} + q(q+1)^{-s-1}\right) + s(s+1)(s+2)(s+3)p^2q^2 2^{-s-4}\\
&\quad -s(s+1)\left(p(p+1)^{-s-2} + q(q+1)^{-s-2} - 3pq - 2^{-s-2}(p^{-s}+q^{-s})\right)\\
&\quad -pqs(s+1)(s+2)\left(2p(p+1)^{-s-3} + 2q(q+1)^{-s-3} - 2^{-s-3}\right)\\
&\quad +s\left(1 - 2^{-s-1} - p^{-s} - q^{-s}\right) - s(s+1)pq2^{-s-2}.
\end{aligned}$$

Theorem 4.1 *For some $\varepsilon > 0$, if $\alpha_1 + \varepsilon \le \frac{k}{\log n} \le \alpha_2 - \varepsilon$ and $t_j = 2\pi j/(\log p/q)$ then*

$$\begin{aligned}
{\sigma_{n,k}^{[X]}}^2 &= G_V^{[X]}\left(\rho_{n,k}, \log_{p/q} p^k n\right) \frac{(p^{-\rho_{n,k}} + q^{-\rho_{n,k}})^k n^{-\rho_{n,k}}}{\sqrt{2\pi\beta(\rho_{n,k})k}}\left(1 + \mathcal{O}(k^{-1/2})\right),\\
{\sigma_{n,k}^{[Y]}}^2 &= G_V^{[Y]}\left(\rho_{n,k}, \log_{p/q} p^k n\right) \frac{(p^{-\rho_{n,k}} + q^{-\rho_{n,k}})^k n^{-\rho_{n,k}}}{\sqrt{2\pi\beta(\rho_{n,k})k}}\left(1 + \mathcal{O}(k^{-1/2})\right),\\
\gamma_{n,k} &= G_C\left(\rho_{n,k}, \log_{p/q} p^k n\right) \frac{(p^{-\rho_{n,k}} + q^{-\rho_{n,k}})^k n^{-\rho_{n,k}}}{\sqrt{2\pi\beta(\rho_{n,k})k}}\left(1 + \mathcal{O}(k^{-1/2})\right),
\end{aligned}$$

where $\rho_{n,k} = \rho(k/\log n) > -2$ satisfies the saddle point equation (9) and

$$\begin{aligned}
G_V^{[X]}(\rho, x) &= \sum_{j\in\mathbb{Z}} \Gamma(\rho + it_j) g_V^{[X]}(\rho + it_j) e^{-2\pi i j x},\\
G_V^{[Y]}(\rho, x) &= \sum_{j\in\mathbb{Z}} \Gamma(\rho + it_j + 1) g_V^{[Y]}(\rho + it_j) e^{-2\pi i j x},\\
G_C(\rho, x) &= \sum_{j\in\mathbb{Z}} \Gamma(\rho + it_j) g_C(\rho + it_j) e^{-2\pi i j x}.
\end{aligned}$$

Tab. 2: Comparisons of magnitudes for $x \in (0,1)$ (Since $\alpha_1 \le \alpha \le \alpha_2$ then $\rho(\alpha) > -2$).

Functions of x	p	$\rho(\alpha)$ -1.5	3.5	8.5	Oscillation
$G_V^{[Y]}(\rho(\alpha), x)$	0.55	0.53	225	1.22×10^{8}	Almost flat
	0.75	0.42	1350	1.58×10^{10}	Non-flat
	0.95	0.13	5×10^{5}	2.50×10^{16}	Non-flat
$G_V^{[X]}(\rho(\alpha), x)$	0.55	0.51	230	1.23×10^{8}	Almost flat
	0.75	0.79	1350	1.58×10^{10}	Non-flat
	0.95	0.85	5×10^{5}	2.50×10^{16}	Non-flat
$G_C(\rho(\alpha), x)$	0.55	-0.05	-220	-1.24×10^{8}	Almost flat
	0.75	-0.15	-1330	-1.50×10^{10}	Non-flat
	0.95	-0.74	-5×10^{5}	-3.50×10^{16}	Non-flat

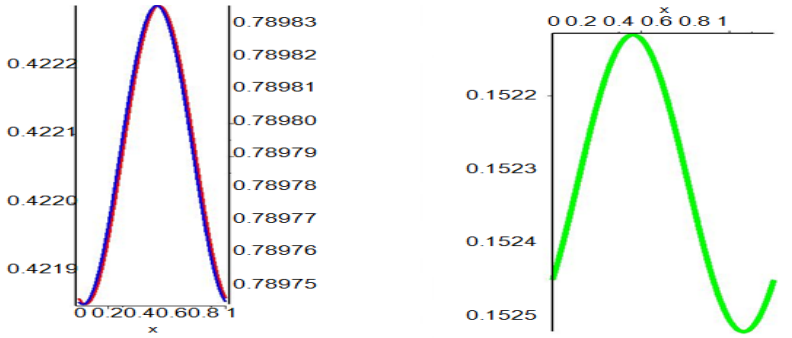


Fig. 3: The fluctuating part of the functions $G_V^{[Y]}(-1.5, x)$ (red curve), $G_V^{[X]}(-1.5, x)$ (blue curve) and $G_C(-1.5, x)$ (green curve), for $x \in (0,1)$ and $p = 0.75$.

Proof: By evaluating the integrals (12), (13) and (14) via the saddle point method, the proof is similar to that of Theorem 3.1 with the new functions $g_V^{[X]}(s)$, $g_V^{[Y]}(s)$ and $g_C(s)$. □

We give examples in Table 2, that the magnitudes of the periodic functions $G_V^{(1)}(\rho(\alpha), x)$, $G_V^{(2)}(\rho(\alpha), x)$ and $G_C(\rho(\alpha), x)$ increase when $\rho(\alpha)$ grows; and also their amplitudes decrease as $p \to 0.5^{+}$, and have oscillation, otherwise. Figure 3 illustrates the oscillation of these functions for $\rho(\alpha) = -1.5$ and $p = 0.75$.

5 Limiting Joint Distribution

In this section, for the second range in Theorem 3.1, *i.e.* $\alpha_0 + \varepsilon \le \frac{k}{\log n} \le \alpha_2 - \varepsilon$ $(-2 < \rho_{n,k} < 0)$, we prove the limiting joint distribution of $X_{n,k}$ and $Y_{n,k}$ is bivariate normal if $\mathbb{V}(X_{n,k}) \to \infty$, $\mathbb{V}(Y_{n,k}) \to \infty$.

Let $F_{n,k}(u, w) := \mathbb{E}[u^{X_{n,k}} w^{Y_{n,k}}]$ denote the joint probability generating function of $X_{n,k}$ and $Y_{n,k}$.

Then

$$F_{n,k}(u,w)=\sum_{l=0}^{n}\binom{n}{j}p^jq^{n-j}F_{j,k-1}(u,w)F_{n-j,k-1}(u,w),\qquad (n\geq 0,k\geq 1),$$

with

$$F_{n,0}(u,w)=\begin{cases}u+n(pq^{n-1}+p^{n-1}q)(w-u), & n\geq 3;\\ u+2pq(w-u), & n=2,\\ 1, & n=0,1.\end{cases}$$

Thus $F_k(x,u,w):=\sum_{n\geq 0}F_{n,k}(u,w)\frac{x^n}{n!}$ satisfies the functional equation

$$F_k(x,u,w)=F_{k-1}(px,u,w)F_{k-1}(qx,u,w),\qquad (k\geq 1),$$

with the initial condition

$$F_0(x,u,w)=e^x+(1-u)(1+x-e^x)+(w-u)(pxe^{qx}+qxe^{px}-x-pqx^2).$$

By iterating this functional equation, we obtain

$$F_k(x,u,w)=\prod_{0\leq j\leq k}F_0(p^jq^{k-j}x,u,w)^{\binom{k}{j}},\qquad (k\geq 1). \tag{15}$$

In the proof of Theorem 5.1, we need the following upper bound for the depoissonization procedure.

Proposition 5.1 *Uniformly for $k\geq 1$, $r\geq 0$, $|\theta|\leq\pi$, $|u|=1$ and $|w|=1$*

$$|F_k(re^{i\theta},u,w)|\leq e^{r-cr\theta^2}, \tag{16}$$

for some constant $c>0$ independent of k, r and θ.

Proof: In order to prove the above upper bound, we need the following second inequality which is hold for $r\geq r_0\approx 2.9183$, $c_1:=2/(3\pi^2)$:

$$(2+r+pqr^2)\big(e^{c_1qr\theta^2/2}+1\big)\leq\left(2+r+\frac{r^2}{4}\right)(e^{r/6}+1)\leq e^r. \tag{17}$$

For $r\leq r_0$, we consider the expansion

$$F_0(x,u,w)=1+x+\frac{x^2}{2}\big(u-2pq(u-w)\big)+\sum_{j\geq 3}\frac{x^j}{j!}\big(u-j(pq^{j-1}+p^{j-1}q)(u-w)\big).$$

Define $\delta_2:=2pq$ and $\delta_j:=j(pq^{j-1}+p^{j-1}q)$, for $j\geq 3$. It is easy to see that $\delta_j\leq 0.75$, for $j\geq 2$ and $0.5\leq p\leq 1$. Thus

$$\begin{aligned}\big|F_0(re^{i\theta},e^{i\varphi},e^{i\psi})\big|&\leq|1+re^{i\theta}|+\sum_{j\geq 2}\frac{r^j}{j!}\big|(1-\delta_j)e^{i\varphi}+\delta_je^{i\psi}\big|\\ &\leq|1+re^{i\theta}|+\sum_{j\geq 2}\frac{r^j}{j!}\\ &\leq e^{r-c_2r\theta^2},\qquad\text{(By (76) in Park et al. (2009))},\end{aligned} \tag{18}$$

uniformly for $1 \le r \le r_0$, $|\theta| \le \pi$ and $c_2 := 2/(\pi^2(1+r_0)^2 e^{r_0})$.

Now suppose $r \ge r_0 \approx 2.9183$. We can rewrite $F_0(x,u,w)$ as follows:

$$F_0(x,u,w) = ua_1(px)a_1(qx) + 1 - u + x + w\left(xqa_2(px) + xpa_2(qx)\right) + wpqx^2,$$

where $a_1(x) := e^x - x$ and $a_2(x) := e^x - 1 - x$. By (17) and applying Lemma 6 in Park et al. (2009),

$$\begin{aligned}\left|F_0(re^{i\theta}, e^{i\varphi}, e^{i\psi})\right| &\le a_1(pr)a_1(qr)e^{-c_1 r\theta^2/2} + qra_2(pr)e^{-c_1 pr\theta^2} + pra_2(qr)e^{-c_1 qr\theta^2} + 2 + r + pqr^2 \\ &\le (e^r+2)e^{-c_1 qr\theta^2} + (2 + r + pqr^2)\left(1 - e^{-c_1 qr\theta^2}\right) \\ &\le e^{r - c_1 qr\theta^2/2}, \qquad \text{(By (17))}. \end{aligned} \tag{19}$$

Collecting the two inequalities (18) and (19), we obtain

$$\left|F_0(re^{i\theta}, e^{i\varphi}, e^{i\psi})\right| \le e^{r - cr\theta^2}, \qquad (c := \min\{c_1 q/2, c_2\}),$$

uniformly for $r \ge 0$, $|\theta| \le \pi$. This implies (16) by (15). □

Now, we prove the following lemma that is needed for the proof of Proposition 5.2.

Lemma 5.1 *The function $Q_k(re^{i\theta}, u, w)$ is well-defined for $r \ge 0$, $|\theta| \le \varepsilon$, $|u| = 1$ and $|w| = 1$.*

Proof: We first show that

$$A := \left|1 - (1 - e^{i\varphi})\big(1 - a_3(r) - a_4(r)\big) - (1 - e^{i\psi})a_4(r)\right| > 0,$$

for $r \ge 0$, $|\theta| \le \varepsilon$, $|u| = 1$ and $|w| = 1$. By direct calculation, we have

$$\begin{aligned} A^2 &= 1 + va_3(r)^2 + za_4(r)^2 + (v + z - t)a_3(r)a_4(r) - va_3(r) - za_4(r) \\ &\ge 1 + va_3(r)^2 + za_4(r)^2 - ta_3(r)a_4(r) - va_3(r) - za_4(r), \end{aligned}$$

where $v := 2(1 - \cos\varphi)$, $t := 2(1 - \cos\psi)$ and $z := 2(1 - \cos(\varphi - \psi))$. Since

$$\begin{aligned} a_4(r) \le \sup_{\substack{r\ge 0 \\ 0.5 \le p \le 1}} a_4(r) &\le \sup_{r \ge 0}\left(pre^{-pr} + qre^{-qr} - re^{-r}\big|_{p=\frac{1}{2}}\right) \\ &= \sup_{r\ge 0} re^{-r}(e^{r/2} - 1) \approx 0.52069, \end{aligned}$$

we have

$$\begin{aligned} A^2 \ge \inf_{\substack{r\ge 0 \\ 0.5\le p\le 1 \\ 0\le v,t,z\le 2}} A^2 &\ge \inf_{\substack{r\ge 0 \\ 0.5\le p\le 1}} \left(1 + va_3(r)^2 + za_4(r)^2 - ta_3(r)a_4(r) - va_3(r) - za_4(r)\Big|_{\substack{v=z=2 \\ t=0}}\right) \\ &= \inf_{\substack{r\ge 0 \\ 0.5\le p\le 1}} \left(1 + 2a_3(r)^2 + 2a_4(r)^2 - 2a_3(r) - 2a_4(r)\right) \\ &= 1 + 2a_3(r)^2 + 2a_4(r)^2 - 2(a_3(r) + a_4(r))\Big|_{\substack{a_3(r)=1 \\ a_4(r)=0.52069}} \approx 0.50085 > 0. \end{aligned}$$

This proves the lemma when $x = r$; the assertion of the lemma follows from analyticity. □

By the same arguments in proofs of Lemma 4 and Theorem 7 in Park et al. (2009), if $-2 < \rho < 0$ or $\alpha_0 \leq \frac{k}{\log n} \leq \alpha_2$, then for $l = 0, 1, 2, \ldots$, we have the following estimates:

$$\begin{aligned}
&\frac{d^l}{dz^l} M_k^{[X]}(z)\Big|_{z=ne^{i\theta}} = \mathcal{O}\Big(n^{-l} M_k^{[X]}(n)\Big), \qquad \frac{d^l}{dz^l} M_k^{[Y]}(z)\Big|_{z=ne^{i\theta}} = \mathcal{O}\Big(n^{-l} M_k^{[Y]}(n)\Big),\\
&\frac{d^l}{dz^l} V_k^{[X]}(z)\Big|_{z=ne^{i\theta}} = \mathcal{O}\Big(n^{-l} V_k^{[X]}(n)\Big), \qquad \frac{d^l}{dz^l} V_k^{[Y]}(z)\Big|_{z=ne^{i\theta}} = \mathcal{O}\Big(n^{-l} V_k^{[Y]}(n)\Big),\\
&\frac{d^l}{dz^l} C_k(z)\Big|_{z=ne^{i\theta}} = \mathcal{O}\Big(n^{-l} C_k(n)\Big), \qquad {\sigma_{n,k}^{[X]}}^2 = \Theta\big(\mu_{n,k}^{[X]}\big), \qquad {\sigma_{n,k}^{[Y]}}^2 = \Theta\big(\mu_{n,k}^{[Y]}\big),\\
&\gamma_{n,k} = \Theta\big({\sigma_{n,k}^{[X]}}^2 + {\sigma_{n,k}^{[Y]}}^2\big), \qquad \gamma_{n,k} \sim C_k(n) - nM_k^{[X]'}(n) M_k^{[Y]'}(n)\\
&\sigma_{n,k}^{[X]} \sim V_k^{[X]}(n) - nM_k^{[X]'}(n)^2, \qquad \sigma_{n,k}^{[Y]} \sim V_k^{[Y]}(n) - nM_k^{[Y]'}(n)^2.
\end{aligned} \tag{20}$$

Proposition 5.2 *Assume that* $\mu_{n,k}^{[X]} \to \infty$ *and* $\mu_{n,k}^{[Y]} \to \infty$. *Then uniformly for* $|\theta| \leq \theta_0 := n^{-2/5}$, $\varphi = o({\sigma_{n,k}^{[Y]}}^{-2/3})$ *and* $\psi = o({\sigma_{n,k}^{[Y]}}^{-2/3})$

$$\begin{aligned}
F_k(ne^{i\theta}, e^{i\varphi}, e^{i\psi}) = \exp\Bigg(&n - \frac{n}{2}\theta^2 + M_k^{[X]}(n) i\varphi + M_k^{[Y]}(n) i\psi - nM_k^{[X]'}(n)\theta\varphi - nM_k^{[Y]'}(n)\theta\psi\\
&-\frac{1}{2}V_k^{[X]}(n)\varphi^2 - \frac{1}{2}V_k^{[Y]}(n)\psi^2 - C_k(n)\varphi\psi + \mathcal{O}(E)\Bigg),
\end{aligned} \tag{21}$$

where

$$\begin{aligned}
E := {}& n|\theta|^3 + {\sigma_{n,k}^{[X]}}^2|\varphi|\theta^2 + {\sigma_{n,k}^{[Y]}}^2\theta^2|\psi| + {\sigma_{n,k}^{[X]}}^2|\theta|\varphi^2 + {\sigma_{n,k}^{[Y]}}^2|\theta|\psi^2\\
&+\gamma_{n,k}|\theta\varphi\psi| + {\sigma_{n,k}^{[Y]}}^2|\psi^3 + \psi\varphi^2 + \varphi\psi^2| + {\sigma_{n,k}^{[X]}}^2|\varphi|^3.
\end{aligned}$$

Proof: Define

$$Q(x, u, w) := \log\big(e^{-x} F_0(x, u, w)\big) = \log\Big(1 - (1-u)\big(1 - a_3(x) - a_4(x)\big) - (1-w)a_4(x)\Big),$$

where $a_3(x) := e^{-x}(1 + x)$ and $a_4(x) := pxe^{-px} + qxe^{-qx} - xe^{-x} - pqx^2e^{-x}$. Let

$$Q_k(x, u, w) := \sum_{j=0}^{k} \binom{k}{j} Q(p^j q^{k-j} x, u, w) = \log\big(e^{-x} F_k(x, u, w)\big).$$

First, we prove in Lemma 5.1 of Appendix that $F_0(re^{i\theta}, e^{i\varphi}, e^{i\psi})$ is away from zero for $r \geq 0$ and $|\theta| \leq \varepsilon$, implying that $Q_k(x, u, w)$ is well-defined when $|\arg(x)| \leq \varepsilon$.

We start from the expansion

$$Q(x,u,w)=\begin{cases}\big((\frac{1}{2}-pq)(1-u)+pq(1-w)\big)x^2+\mathcal{O}\big(|2-u-w||x|^3\big), & \text{as } x\to 0;\\ (1-u)\big(1+\mathcal{O}(|x|e^{-q\Re(x)})\big), & \text{as } x\to\infty, |\arg(x)|\le\varepsilon.\end{cases}$$

By the above expansion, we have

$$Q_k(x,u,w)=\frac{1}{2\pi i}\int_{\rho-i\infty}^{\rho+i\infty}x^{-s}Q^*(s,u,w)(p^{-s}+q^{-s})^k\mathrm{d}s,$$

where $-2<\rho<0$ and $Q^*(s,u,w):=\int_0^\infty x^{s-1}Q(x,u,w)dx$ is defined for $-2<\Re(s)<0$. Note that

$$\begin{aligned}Q(x,u,w)&=(1-u)\big(1-a_3(x)-a_4(x)\big)+(1-w)a_4(x)-\frac{1}{2}\Big((1-u)^2\big(1-a_3(x)-a_4(x)\big)^2\\&\quad+2(1-u)(1-w)\big(1-a_3(x)-a_4(x)\big)a_4(x)+(1-w)^2a_4(x)^2\Big)\\&\quad+\hat{Q}(x,u,w)(1-u)^3+\tilde{Q}(x,u,w)(1-u)^2(1-w)\\&\quad+\breve{Q}(x,u,w)(1-u)(1-w)^2+\bar{Q}(x,u,w)(1-w)^3,\end{aligned}$$

where the exact forms of $\hat{Q}$, $\tilde{Q}$, $\breve{Q}$ and $\bar{Q}$ can be obtained by Taylor's reminder formula and are of less important here. We need instead the estimates (the assumptions in Lemma 8 in Park et al. (2009))

$$\mathcal{O}(\hat{Q}(x,u,w))=\mathcal{O}(\breve{Q}(x,u,w))=\mathcal{O}(\tilde{Q}(x,u,w))=\mathcal{O}(\bar{Q}(x,u,w))=\mathcal{O}\big(|x|^6\big)=\mathcal{O}\big(|x|^2\big),$$

as $x\to 0$ and

$$\begin{aligned}\tilde{Q}(x,u,w)&=\mathcal{O}(|x|e^{-q\Re(x)}),\\ \breve{Q}(x,u,w)&=\mathcal{O}(|x|^2e^{-2q\Re(x)})=\mathcal{O}(|x|e^{-q\Re(x)}),\\ \bar{Q}(x,u,w)&=\mathcal{O}(|x|^3e^{-3q\Re(x)})=\mathcal{O}(|x|e^{-q\Re(x)}),\\ \hat{Q}(x,u,w)&=1+\mathcal{O}(|x|e^{-q\Re(x)}),\end{aligned}\tag{22}$$

as $x\to\infty$ in the sector $\{x:|\arg(x)|\le\varepsilon\}$. By (5), (6), (10) and (11), this expansion gives

$$\begin{aligned}Q_k(x,u,w)&=(1-u)M_k^{[X]}(x)+(1-w)M_k^{[Y]}(x)+\frac{1}{2}(1-u)^2\big(V_k^{[X]}(x)-M_k^{[X]}(x)\big)\\&\quad+\frac{1}{2}(1-w)^2\big(V_k^{[Y]}(x)-M_k^{[Y]}(x)\big)+(1-u)(1-w)C_k(x)\\&\quad+(1-u)^3\hat{Q}_k(x,u,w)+(1-u)^2(1-w)\tilde{Q}_k(x,u,w)\\&\quad+(1-u)(1-w)^2\breve{Q}_k(x,u,w)+(1-w)^3\bar{Q}_k(x,u,w),\end{aligned}$$

where $\tilde{Q}$, $\breve{Q}$ and $\bar{Q}$ satisfy in (6); and $\hat{Q}$ satisfy in (5).

Applying Lemma 8 in Park et al. (2009) and expansions in (22), we have $\tilde{Q}_k(x,u,w)=\Theta(M_k^{[Y]}(x))$, $\breve{Q}_k(x,u,w)=\Theta(M_k^{[Y]}(x))$ and $\bar{Q}_k(x,u,w)=\Theta(M_k^{[Y]}(x))$; and similarly $\hat{Q}_k(x,u,w)=\Theta(M_k^{[X]}(x))$.

These estimates yield, with $x = ne^{i\theta}$,

$$\begin{aligned}Q_k(x,u,w) &= (1-u)M_k^{[X]}(x) + (1-w)M_k^{[Y]}(x) + \frac{1}{2}(1-u)^2\big(V_k^{[X]}(x) - M_k^{[X]}(x)\big)\\ &+\frac{1}{2}(1-w)^2\big(V_k^{[Y]}(x) - M_k^{[Y]}(x)\big) + (1-u)(1-w)C_k(x)\\ &+\mathcal{O}\Big(\big|1-u\big|^3\big|M_k^{[X]}(ne^{i\theta})\big|\Big) + \mathcal{O}\Big(\big|(2-u-w)^3-(1-u)^3\big|\big|M_k^{[Y]}(ne^{i\theta})\big|\Big),\end{aligned}$$

where the $\mathcal{O}$-term holds uniformly for $|\theta| \le \varepsilon$ and $|1-u| = o(1)$ and $|(2-u-w)^3-(1-u)^3| = o(1)$. Since $\mu_{n,k}^{[X]} \to \infty$ and $\mu_{n,k}^{[Y]} \to \infty$, this leads to (21) by expansions of $M_k^{[X]}(ne^{i\theta})$, $M_k^{[Y]}(ne^{i\theta})$, $V_k^{[X]}(ne^{i\theta})$, $V_k^{[Y]}(ne^{i\theta})$ and $C_k(ne^{i\theta})$ at $\theta = 0$, using the estimates in (20). □

Theorem 5.1 *For $\alpha_0 + \varepsilon \le \frac{k}{\log n} \le \alpha_2 - \varepsilon$, if $\sigma_{n,k}^{[X]^2} \to \infty$ and $\sigma_{n,k}^{[Y]^2} \to \infty$ then*

$$\mathbb{P}\left(\frac{X_{n,k}-\mu_{n,k}^{[X]}}{\sigma_{n,k}^{[X]}} \le x, \frac{Y_{n,k}-\mu_{n,k}^{[Y]}}{\sigma_{n,k}^{[Y]}} \le y\right) = \Phi(x,y;\rho_{n,k}) + o(1),$$

where $\rho_{n,k} := \gamma_{n,k}/\sigma_{n,k}^{[X]}\sigma_{n,k}^{[Y]}$ and $\Phi(x,y;\rho)$ denotes the cumulative distribution function of bivariate standard normal distribution with correlation parameter ρ.

Proof: Recall that $\theta_0 := n^{-2/5}$. By Cauchy's integral formula, (16) and (21), we have

$$\begin{aligned}\mathbb{E}\left(e^{X_{n,k}i\varphi+Y_{n,k}i\psi}\right) &= \frac{n!}{2\pi i}\int_{|x|=n} x^{-n-1}F_k(x,e^{i\varphi},e^{i\psi})\mathrm{d}x\\ &= \frac{n!n^{-n}}{2\pi}\int_{|\theta|\le\theta_0} x^{-n-1}F_k(ne^{i\theta},e^{i\varphi},e^{i\psi})\mathrm{d}\theta + \mathcal{O}\left(n^{-1/10}e^{-cn^{1/5}}\right)\\ &= \frac{n!n^{-n}}{2\pi}e^{n+M_k^{[X]}(n)i\varphi+M_k^{[Y]}(n)i\psi-\frac{1}{2}V_k^{[X]}(n)\varphi^2-\frac{1}{2}V_k^{[Y]}(n)\psi^2-C_k(n)\varphi\psi}\\ &\times\int_{-\theta_0}^{\theta_0} e^{-\frac{n}{2}\theta^2-nM_k^{[X]\prime}(n)\theta\varphi-nM_k^{[Y]\prime}(n)\theta\psi}\big(1+\mathcal{O}(E)\big)\mathrm{d}\theta + \mathcal{O}\left(n^{-1/10}e^{-cn^{1/5}}\right),\end{aligned}$$

since $E \to 0$ in the range of integration and when $\varphi = o(\sigma_{n,k}^{[X]^{-4/5}})$ and $\psi = o(\sigma_{n,k}^{[Y]^{-4/5}})$. Applying Stirling's formula, extending the integration limits to $\pm\infty$ and making the change of variables $\theta \mapsto \theta n^{-1/2}$, uniformly in φ and ψ, we obtain

$$\begin{aligned}\mathbb{E}\left(e^{X_{n,k}i\varphi+Y_{n,k}i\psi}\right) &= \frac{1}{\sqrt{2\pi}}e^{M_k^{[X]}(n)i\varphi+M_k^{[Y]}(n)i\psi-\frac{\varphi^2}{2}\left(V_k^{[X]}(n)-nM_k^{[X]\prime}(n)^2\right)-\frac{\psi^2}{2}\left(V_k^{[Y]}(n)-nM_k^{[Y]\prime}(n)^2\right)}\\ &\times e^{-\varphi\psi\left(C_k(n)-nM_k^{[X]\prime}(n)M_k^{[Y]\prime}(n)\right)}\times\int_{-\infty}^{\infty} e^{-\left(\theta+\sqrt{n}M_k^{[X]\prime}(n)\varphi+\sqrt{n}M_k^{[Y]\prime}(n)\psi\right)^2/2}\\ &\times\Bigg(1+\mathcal{O}\Bigg(\frac{1+|\theta|^3}{\sqrt{n}}+\frac{\theta^2}{n}\left(\sigma_{n,k}^{[X]^2}|\varphi|+\sigma_{n,k}^{[Y]^2}|\psi|\right)+\frac{|\theta|}{\sqrt{n}}\left(\sigma_{n,k}^{[X]^2}\varphi^2+\sigma_{n,k}^{[Y]^2}\psi^2\right.\\ &\qquad\left.+\gamma_{n,k}|\varphi\psi|\right)+\sigma_{n,k}^{[Y]^2}|\psi^3+\psi\varphi^2+\varphi\psi^2|+\sigma_{n,k}^{[X]^2}|\varphi|^3\Bigg)\Bigg)\mathrm{d}\theta,\end{aligned}$$

$$\longrightarrow \exp\Big(\mu_{n,k}^{[X]} i\varphi + \mu_{n,k}^{[Y]} i\psi - \frac{\varphi^2}{2}{\sigma_{n,k}^{[X]}}^2 - \frac{\psi^2}{2}{\sigma_{n,k}^{[Y]}}^2 - \varphi\psi\gamma_{n,k}\Big)$$
$$\times\Big(1 + \mathcal{O}\Big({\sigma_{n,k}^{[Y]}}^2|\psi^3 + \psi\varphi^2 + \varphi\psi^2| + {\sigma_{n,k}^{[X]}}^2|\varphi|^3\Big)\Big), \qquad \text{(by (20))},$$

which implies the result by Lévy's continuity theorem. □

6 Main Results for Symmetric Tries

When $p = q = 1/2$, the major difference is reflected by the fact that $\alpha_1 = \alpha_2$, so that the saddle point range between α_1 and α_2 does not exist, and most analysis we give above becomes much simpler. For simplicity of presentation, we omit all error terms in our asymptotic estimates

Asymptotics of the expectations. From (5), we have

$$M_k^{[X]}(x) = 2^k - 2^k e^{-x/2^k} - xe^{-x/2^{k+1}} + 2^{-k-2}x^2e^{-x/2^k}, \qquad (k \geq 0).$$

By this and de-Poissonization procedures (Proposition 1 and Lemma 4 in Park et al. (2009)), we deduce

$$\mathbb{E}(X_{n,k}) \sim \begin{cases} 2^k - n(1-2^{-k-1})^{n-1}, & \text{if } 2^{-k}n \to \infty; \\ M_k^{[X]}(n), & \text{if } 4^{-k}n \to 0. \end{cases}$$

In particular,

$$\mathbb{E}(X_{n,k}) \sim \begin{cases} 2^k(1 + 2^{-2}t^2e^{-t} - e^{-t} - te^{-t/2}), & \text{if } 2^{-k}n \to t \in (0,\infty); \\ 2^{-k-2}n^2, & \text{if } 2^{-k}n \to 0. \end{cases}$$

In a similar manner, we have, by (6),

$$M_k^{[Y]}(x) = xe^{-x/2^{k+1}} - xe^{-x/2^k} - 2^{-k-2}x^2e^{-x/2^k}, \qquad (k \geq 0).$$

Therefore, we have

$$\mathbb{E}(Y_{n,k}) \sim \begin{cases} n(1-2^{-k-1})^{n-1}, & \text{if } 2^{-k}n \to \infty; \\ M_k^{[Y]}(n), & \text{if } 4^{-k}n \to 0. \end{cases}$$

This implies that

$$\mathbb{E}(Y_{n,k}) \sim \begin{cases} n(e^{-t/2} - e^{-t} - 2^{-2}te^{-t}), & \text{if } 2^{-k}n \to t \in (0,\infty); \\ 2^{-k-2}n^2, & \text{if } 2^{-k}n \to 0. \end{cases}$$

Asymptotics of the variances. Similarly, by (10), we have

$$V_k^{[X]}(x) = 2^k - 2^k e^{-x/2^k} - xe^{-x/2^{k+1}} + 2^{-k-2}x^2e^{-x/2^k}$$
$$- 2^k\big(1 - e^{-x/2^k} - 2^{-k}xe^{-x/2^{k+1}} + 2^{-2k-2}x^2e^{-x/2^k}\big)^2,$$
$$V_k^{[Y]}(x) = xe^{-x/2^{k+1}} - xe^{-x/2^k} - 2^{-k-2}x^2e^{-x/2^k}$$
$$- 2^{-k}\big(xe^{-x/2^{k+1}} - xe^{-x/2^k} - 2^{-k-2}x^2e^{-x/2^k}\big)^2,$$

and, if $n/2^k \to \infty$, then

$$\mathbb{V}(X_{n,k}) \sim \mathbb{V}(Y_{n,k}) \sim \mathbb{E}(Y_{n,k}) \sim n(1-2^{-k-1})^{n-1};$$

and if $n/4^k \to 0$, then

$$\mathbb{V}(X_{n,k}) \sim V_k^{[X]}(x), \quad \text{and} \quad \mathbb{V}(Y_{n,k}) \sim V_k^{[Y]}(x),$$

uniformly in k. These approximations imply that

$$\mathbb{V}(X_{n,k}) \sim \begin{cases} 2^k(1-e^{-t}-te^{-t/2}+2^{-2}t^2e^{-t})(e^{-t}+te^{-t/2}-2^{-2}t^2e^{-t}), & \text{if } 2^{-k}n \to t \in (0,\infty); \\ 2^{-k-2}n^2, & \text{if } 2^{-k}n \to 0. \end{cases}$$

and

$$\mathbb{V}(Y_{n,k}) \sim \begin{cases} n\big(e^{-t/2}-e^{-t}-2^{-2}te^{-t}\big)\big(1-t(e^{-t/2}-e^{-t}-2^{-2}te^{-t})\big), & \text{if } 2^{-k}n \to t \in (0,\infty); \\ 2^{-k-2}n^2, & \text{if } 2^{-k}n \to 0. \end{cases}$$

Limiting joint distribution Theorem 5.1 holds when $p = q = 1/2$ by the same method of proof. Note that the trivariate generating function becomes simpler (see (15))

$$F_k(x,u,w) = \left(e^{x/2^k} + (1-u)\Big(1+\frac{x}{2^k}-e^{x/2^k}\Big) + (w-u)\Big(\frac{x}{2^k}e^{x/2^k} + \frac{x}{2^k}e^{x/2^k} - \frac{x}{2^k} - \frac{x^2}{4^{k+1}}\Big)\right)^{2^k}.$$

References

G.-S. Cheon and L. W. Shapiro. Protected points in ordered trees. *Appl. Math. Lett.*, 21:516–520, 2008.

L. Devroye and S. Janson. Protected nodes and fringe subtrees in some random trees. *Electron. Commun. Probab*, 19:1–10, 2014.

M. Drmota. *Random trees: An interplay between combinatorics and probability*. Springer Wien New York, Vienna, 2009.

R. R. Du and H. Prodinger. Notes on protected nodes in digital search trees. *Appl. Math. Lett.*, 25: 1025–1028, 2012.

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

M. Fuchs and C.-K. Lee. A general central limit theorem for shape parameters of m-ary tries and patricia tries. 21:1–26, 2014.

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

M. Fuchs, C.-K. Lee, and G.-R. Yuctiker. On 2-protected nodes in random digital trees. *Theor. Comput. Sci.*, 622:111–122, 2016.

J. Gaither and M. D. Ward. The variance of the number of 2-protected nodes in a trie. In *The Tenth Workshop on Analytic Algorithmics and Combinatorics*, page 43–51, 2013.

J. Gaither, Y. Homma, M. Sellke, and M. D. Ward. On the number of 2-protected nodes in tries and suffix trees. In *Discrete Math. Theor. Comput. Sci. Proc.*, page 381–398, 2012.

H.-K. Hwang, M. Fuchs, and V. Zacharovas. Asymptotic variance of random symmetric digital search trees. *Discrete Math. Theor. Comput. Sci.*, 12:103–166, 2012.

D. Knuth. *The art of computer programming. Sorting and searching*. Addison-Wesley, Reading, MA, New York, 1998.

H. M. Mahmoud. *Evolution of random search trees*. John Wiley & Sons Inc., New York, 1992.

P. Nicodème. Average profiles, from tries to suffix-trees. In *Discrete Math. Theor. Comput. Sci. Proc.*, page 257–266, 2005.

G. Park, H.-K. Hwang, P. Nicodème, and W. Szpankowski. Profiles of tries. *SIAM J. Computing*, 38: 1821–1880, 2009.

W. Szpankowski. *Average case analysis of algorithms on sequences*. Wiley-Interscience, New York, 2001.