On 2-Protected Nodes in Random Digital Trees

M. Fuchs\textsuperscript{a}, C.-K. Lee\textsuperscript{a,}\textsuperscript{*}, G.-R. Yu\textsuperscript{a}

\textsuperscript{a}Department of Applied Mathematics, National Chiao Tung University, Hsinchu, 300, Taiwan

Abstract

In this paper, we consider the number of 2-protected nodes in random digital trees. For tries, Gaither, Homma, Sellke and Ward (2012) and Gaither and Ward (2013) derived asymptotic expansions for mean and variance. We re-derive their results with a different approach. The expressions we obtain are quite different and contain divergent series which have values by appealing to the theory of Abel summability. For PATRICIA tries, we prove similar results. Moreover, we prove a bivariate central limit theorem for the size and the number of 2-protected nodes in tries. This result entails that the numbers of 2-protected nodes in tries and PATRICIA tries both satisfy a central limit theorem. Finally, we derive similar results for 2-protected nodes in digital search trees as well.

Keywords: 2-protected nodes, digital trees, moments, central limit theorems, Abel summability

2010 MSC: 60C05; 60F05; 68P05

1. Introduction

Several recent studies have been concerned with so-called 2-protected nodes which are nodes in trees with a distance of at least two from any leaf. They have been, e.g., studied for random binary search trees [1, 2, 11], random ternary search trees [8], random recursive trees [2, 12] and simple generated families of random trees [2].

In this paper, we are interested in the number of 2-protected nodes in the three main families of random digital trees, namely, random tries, random PATRICIA tries and random digital search trees. These three families are of fundamental importance in computer science; for background and precise definition see for instance [10] or [4, 9].

We first recall what is known about the number of 2-protected nodes in random digital trees. The first paper which studied this parameter was [3] where an asymptotic expansion of the mean in symmetric digital search trees was derived. Moreover, a similar result was proved for tries in [7]. In addition, the authors in [6] found an asymptotic expansion for the variance of the number of 2-protected nodes in tries and conjectured a central limit theorem.

The aim of this paper is to show that the tools of [4] and [5] can applied to the number of 2-protected nodes in random digital trees. More precisely, we will re-derive the results from [6, 7] and add similar results for PATRICIA tries. Expressions of periodic functions in our results will considerably differ from [6, 7] and will contain divergent series which can be made convergent by appealing to the theory of Abel summability. Apart from asymptotic expansions of moments, we will also prove a bivariate central limit theorem for the number of 2-protected nodes and the size of random tries. As a consequence of this result, we obtain the conjecture from [6]. In addition, this result also yields a central limit theorem for PATRICIA tries. Finally, we will prove similar results, namely, asymptotic expansions for mean and variance and a central limit theorem also for symmetric digital search trees.

Notations. Throughout the paper, the number of 2-protected nodes in a random digital tree of size \( n \) under the Bernoulli model will be denoted by \( X_{n}^{(\ast)} \) with \( \ast \in \{ T, P, D \} \), depending on whether tries, PATRICIA tries, or digital search trees are considered. The probability that a bit equals 1 in the Bernoulli model will be \( p \in (0, 1) \) and we set \( q := 1 - p \). Finally, for some function \( G(x) \), we will use the notation

\[
\mathcal{F}[G](x) := \begin{cases} 
\frac{1}{h} \sum_{k \in \mathbb{Z} \setminus \{0\}} G(-1 + \chi_{k}) e^{2k\pi i x}, & \text{if } \frac{\log p}{\log q} \in \mathbb{Q}; \\
0, & \text{if } \frac{\log p}{\log q} \not\in \mathbb{Q},
\end{cases}
\]

*Corresponding author: Email: enix45@gmail.com; Phone: +88635712121 Ext. 56461
where \( h = -p \log p - q \log q \) and \( \chi_k = 2r k \pi i / \log p \) when \( \log p / \log q = r/l \) with \( \gcd(r, l) = 1 \).

2. 2-Protected Nodes in Tries and PATRICIA tries

**Distributional Recurrences.** We start with tries. The main observation is that the number of 2-protected nodes can be computed recursively by computing the number for the two subtrees and then adding 1 if the root itself is 2-protected. The latter happens if and only if neither the left nor the right subtree contains only one data. This leads to the following distributional recurrence for \( X_n^{(T)} \)

\[
X_n^{(T)} = \begin{cases} 
X_{n-1}^{(T)}, & \text{if } I_n = \{1, n-1\}; \\
X_{n-1}^{(T)} + X_{n-I_n}^{(T)} + 1, & \text{otherwise,}
\end{cases} \quad (n \geq 2)
\]

with initial conditions \( X_0^{(T)} = X_1^{(T)} = 0 \). Here, \( X_n^{(T)*} \) denotes an independent copy of \( X_n^{(T)} \) and \( I_n \) is the size of the left subtree which under the Bernoulli model has the distribution

\[
P(I_n = k) = \binom{n}{k} p^k q^{n-k}; \quad (0 \leq k \leq n).
\]

Similarly, for PATRICIA tries, one obtains the slightly different recurrence

\[
X_n^{(P)} = \begin{cases} 
X_{n-1}^{(P)}, & \text{if } I_n = \{0, n\}; \\
X_{n-1}^{(P)} + X_{n-I_n}^{(P)} + 1, & \text{otherwise.}
\end{cases} \quad (n \geq 2),
\]

where notation is as above and initial conditions are again \( X_0^{(P)} = X_1^{(P)} = 0 \).

Note that the number of 2-protected nodes in tries and PATRICIA tries are connected as follows via the number \( N_n \) of internal nodes in tries

\[
X_n^{(P)} = X_n^{(T)} - N_n + n - 1. \quad (1)
\]

The reason for this is that a PATRICIA trie differs from the trie by the nodes with one-way branching which are counted by \( N_n - n + 1 \) and all these nodes are 2-protected.

**Moments.** Due to the above distributional recurrences, 2-protected nodes are additive shape parameters. In [4] a general framework for deriving asymptotic expansions of mean and variance of additive shape parameters in random tries and random PATRICIA tries was given. The framework can be applied to 2-protected nodes and yields the following result.

**Theorem 1** (Mean of 2-protected nodes in tries). The mean of the number of 2-protected nodes in tries satisfies

\[
\frac{\mathbb{E}(X_n^{(T)})}{n} = \frac{pq + 1 - h}{h} + \mathcal{O}(\log_{1/p} n) + o(1),
\]

where for \( \log p / \log q \in \mathbb{Q} \) and \( k \neq 0 \),

\[
G_1^{(T)}(-1 + \chi_k) = \Gamma(-1 + \chi_k) \chi_k (\chi_k pq - pq - 1).
\]

**Remark 1.** This result coincides with the result from [7]. Note, however, that no periodic function was given in [7] and the error term was wrongly stated as \( \mathcal{O}(1/n) \).

For the variance, again the framework from [4] can be applied, but this time some more computations are necessary. We will only state the result without giving details, but will offer several remarks. We need the following function \( K_1(s) = -1 + pq s + s - p^{-s} s - q^{-s} s \).

**Theorem 2** (Variance of 2-protected nodes in tries). The variance of the number of 2-protected nodes in tries satisfies

\[
\frac{\mathbb{V}(X_n^{(T)})}{n} = \frac{G_2^{(T)}(-1)}{h} + \mathcal{O}(\log_{1/p} n) + o(1),
\]

where \( G_2^{(T)}(-1 + \chi_k) \) for \( \log p / \log q \in \mathbb{Q} \) and \( k \neq 0 \) is given by
\[-\Gamma(\chi_k + 3)2^{-3-\chi_k}p^2q^2 + \Gamma(\chi_k + 2)(2p^2q(1 + p)^{-2-\chi_k} + 2pq^2(1 + q)^{-2-\chi_k}) + \Gamma(\chi_k + 1)(-3pq + 2^{-\chi_k}pq - 2^{-1-\chi_k}) + \Gamma(\chi_k)(1 - 2p(1 + p)^{-\chi_k} - 2q(1 + q)^{-\chi_k}) + \Gamma(\chi_k - 1)(1 - 2^{1-\chi_k}) + 2 \sum_{\ell \geq 2} \frac{(-1)^\ell}{\ell!} \frac{K_1(-\ell)}{1 - p^\ell - q^\ell} \left(p^\ell + q^\ell\right) (\ell - 1)\Gamma(\chi_k + \ell + 1) + (1 - \ell \left(p^\ell + q^\ell\right) (2pq + p^{-\ell+2} + q^{-\ell+2})) \Gamma(\chi_k + \ell) + (p^\ell + q^\ell) (1 - \ell + p^{-\ell+2} + q^{-\ell+2}) \Gamma(\chi_k + \ell - 1) + 2 \sum_{\ell \geq 1} \frac{(-1)^\ell}{\ell!} \frac{p^{1+\ell} + q^{1+\ell}}{1 - p^{1+\ell} - q^{1+\ell}} K_1(\chi_k + \ell - 1)K_1(\ell - 1)\Gamma(\chi_k + \ell) + \frac{1}{h} \sum_{j \in \mathbb{Z}} (\chi_j - 1)G_{1}^{(T)}(\chi_j - 1)(\chi_k - j - 1)G_{1}^{(T)}(\chi_k - j - 1)\]

and $G_{2}^{(T)}(-1)$ is given by

\[
\frac{2p^2q}{(1 + p)^2} + \frac{2pq^2}{(1 + q)^2} - 2pq - \frac{p^2q^2}{4} + 2p \log(1 + p) + 2q \log(1 + q) + \frac{1}{2} + h - 2 \log 2 + 2 \sum_{\ell \geq 2} \frac{(-1)^\ell}{\ell!} \frac{K_1(-\ell)}{1 - p^\ell - q^\ell} \left(p^\ell + q^\ell\right) (\ell + 1)\Gamma(\chi_k + \ell + 1) + 2 \sum_{\ell \geq 1} \frac{(-1)^\ell}{\ell!} \frac{p^{1+\ell} + q^{1+\ell}}{1 - p^{1+\ell} - q^{1+\ell}} K_1(\ell - 1)K_1(-\ell - 1) + \frac{1}{h} (pq + 1 - h)^2 \left(p^\ell + q^\ell\right) (\ell + 1)\Gamma(\chi_k + \ell + 1) - \left\{ \begin{array}{ll}
\frac{1}{h \log p} \sum_{j \geq 1} \frac{4rj\pi^2}{\sinh\left(\frac{2rj\pi^2}{\log p}\right)} \left(\frac{2rj\pi p}{\log p}\right)^2 + (pq + 1)^2, & \text{if } \frac{\log p}{\log q} \in \mathbb{Q}; \\
0, & \text{if } \frac{\log p}{\log q} \notin \mathbb{Q}.
\end{array} \right.
\]

Remark 2. Comparing with the result from [6] note that our expressions above are quite different (the authors in [6] forgot the last term in $G_{2}^{(T)}(-1)$ and also stated the error term wrongly as $O(n^{-e})$). It is, however, not complicated to show that our expression coincides with that from [6]. For fixed $p$ equality can also be shown numerically, e.g., for the symmetric case, we obtain

\[
\frac{G_{2}^{(T)}(-1)}{\log 2} \approx 0.93443870447019249853 \cdots
\]

which is the same as the value given in [6].

Remark 3. In contrast to [6], the series expressions in the above result are not convergent in the classical sense. However, they do converge (and give the correct value) if one uses Abel summability. In order to give an explanation for this consider the third last term of $G_{2}^{(T)}(-1)$ in the symmetric case which is obtained by moving the line of integration to $+\infty$ and applying the residue theorem to

\[
\frac{1}{2\pi i} \int_{(0)^+} \frac{2^{-\omega}}{1 - 2^{-\omega}} (1 - \omega)\Gamma(\omega - 1)K_1(\omega - 1)(1 + \omega)\Gamma(-1 - \omega)K_1(-1 - \omega) d\omega,
\]

where integration is along the line $\Re(\omega) = 0$ with a small indentation to the right at zeros of $1 - 2^{-\omega}$. The other series expressions are obtained similarly.

For the sake of simplicity, we consider the following simplified integral (the argument will be the same for the above integral and all similar integrals)

\[
\frac{1}{2\pi i} \int_{(0)^+} \frac{\Gamma(-\omega + 1)\Gamma(\omega)}{1 - 2^{-\omega}} d\omega.
\]
Observe that moving the line of integration to $+\infty$ gives the following divergent series
\[
\sum_{\ell \geq 1} \frac{(-1)^{\ell-1}}{1 - 2^{-\ell}}.
\] (3)

If on the other hand, we consider Abel summability, then (2) has to be replaced by
\[
\frac{1}{2\pi i} \int_{(0)^+} \frac{\Gamma(-\omega + 1)\Gamma(\omega)}{1 - 2^{-\omega}} x^\omega \, d\omega
\]
with $|x| < 1$. Now, moving the line of integration to $+\infty$ is possibly since $x^\omega$ decays exponentially fast to zero. This yields
\[
\sum_{\ell \geq 1} \frac{(-1)^{\ell-1}}{1 - 2^{-\ell}} x^\ell.
\]

Thus, we have the identity
\[
\frac{1}{2\pi i} \int_{(0)^+} \frac{\Gamma(-\omega + 1)\Gamma(\omega)}{1 - 2^{-\omega}} x^\omega \, d\omega = \sum_{\ell \geq 1} \frac{(-1)^{\ell-1}}{1 - 2^{-\ell}} x^\ell.
\]

Finally, letting $x$ tend to 1 yields (2) on the left-hand side and the Abel sum of (3) on the right-hand side. This shows our claim.

Note that one alternatively could move the line of integration of (2) to $-\infty$ which would give a convergent series, however, such an evaluation does not work in the asymmetric case (but this could be used to give an alternative expression of $G_2^{(T)}(-1 + \chi_k)$ for $p = q = 1/2$; see [4] where several such expressions were given, e.g., for the size of tries).

We next turn to PATRICIA tries for which similar results can be proved as above (in order to keep the paper short, we do not give the expressions for the Fourier coefficients of the periodic functions which are similar as above).

**Theorem 3** (Mean and variance of 2-protected nodes in PATRICIA tries). The mean of the number of 2-protected nodes in PATRICIA tries is given by
\[
\frac{\mathbb{E}(X_n^{(P)})}{n} = \frac{pq}{h} + \mathcal{F}[G_1^{(P)}](r \log_1/p \cdot n) + o(1)
\]
and the variance is given by
\[
\frac{\text{Var}(X_n^{(P)})}{n} = \frac{G_2^{(P)}(-1)}{h} + \mathcal{F}[G_2^{(P)}](r \log_1/p \cdot n) + o(1),
\]
where $G_1^{(P)}(x)$ and $G_2^{(P)}(x)$ are two computable functions.

**Remark 4.** Note that the result for the mean either follows from the framework in [4] or alternatively from (1) and known results for $N_n$ (e.g., see [4]).

**Remark 5.** From (1), we obtain that
\[
\text{Cov}(N_n, X_n^{(T)}) = \frac{1}{2} \left( \text{Var}(X_n^{(T)}) + \text{Var}(N_n) - \text{Var}(X_n^{(P)}) \right).
\]

Thus, the above results for the variances of $X_n^{(T)}$ and $X_n^{(P)}$ and the known result for the variance of $N_n$ (e.g., see [4]) gives the following result.

**Theorem 4** (Covariance of size and 2-protected nodes in tries). The covariance of the number of internal nodes and the number of 2-protected nodes in tries satisfies
\[
\frac{\text{Cov}(N_n, X_n^{(T)})}{n} = \frac{H_2(-1)}{h} + \mathcal{F}[H_2](r \log_1/p \cdot n) + o(1),
\]
where
\[
H_2(x) = \frac{G_2^{(T)}(x) + G_2^{(N)}(x) - G_2^{(P)}(x)}{2}
\]
with $G_2^{(N)}(x)$ is given in [4].
Limit Laws. Here, we use the approach from [5] to show a bivariate central limit theorem for $N_n$ and $X_n^{(T)}$. First, we set

$$
\Sigma_n = n \begin{pmatrix}
G_2(N)(-1)/h + \mathcal{F}[G_2(N)](r \log_{1/p} n) & H_2(-1)/h + \mathcal{F}[H_2](r \log_{1/p} n) \\
H_2(-1)/h + \mathcal{F}[H_2](r \log_{1/p} n) & G_2(T)(-1)/h + \mathcal{F}[G_2(T)](r \log_{1/p} n)
\end{pmatrix},
$$

For normalization purpose, we need to show that $\Sigma_n$ is positive definite for large $n$.

**Lemma 1.** For all $n$ large enough, we have that $\Sigma_n$ is positive definite.

**Proof.** It suffices to show that $\text{Var}(N_n) \geq cN n$ and $\det \Sigma_n > 0$ for $n$ large enough. The first claim is classical and the second follows with a similar method of proof as Proposition 3 in [5].

Thus, we can consider $\Sigma_n^{-1/2}$ if $n$ is large enough. Our main result in this section is the following bivariate central limit theorem.

**Theorem 5.** We have,

$$
\Sigma_n^{-1/2} \begin{pmatrix}
N_n - \mathbb{E}(N_n) \\
X_n^{(T)} - \mathbb{E}(X_n^{(T)})
\end{pmatrix} \xrightarrow{d} N(0, I_2),
$$

where $I_2$ denotes the $2 \times 2$ unity matrix and $N(0, I_2)$ is the standard two-dimensional normal distribution.

**Proof.** This follows from our expressions for mean and variance of $X_n^{(T)}$ with a similar method of proof as for Theorem 4 in [5].

As a consequence of this result, we have the following corollary.

**Corollary 1.** For $* \in \{T, P\}$, we have

$$
\frac{X_n^{(*)} - \mathbb{E}(X_n^{(*)})}{\sqrt{\text{Var}(X_n^{(*)})}} \xrightarrow{d} N(0, 1).
$$

3. 2-Protected Nodes in Digital Search Trees

In [4, 5], we proposed general frameworks for deriving asymptotic expansions of moments and central limit theorems for additive shape parameters in tries. In fact, using the tools from [9], similar frameworks can be given for additive shape parameter in random symmetric digital search trees as well. These frameworks can then be applied to the number of 2-protected nodes in digital search trees which is easily seen to be an additive shape parameter. We only give the result here.

**Theorem 6** (2-protected nodes in symmetric digital search trees). The mean and the variance of the number of 2-protected nodes in symmetric digital search trees satisfy

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

and

$$
\text{Var}(X_n^{(D)}) = \frac{n}{\log 2} \sum_{k \in \mathbb{Z}} \frac{G_2^{(D)}(2 + \chi_k)}{\Gamma(2 + \chi_k)} + \mathcal{O}(1),
$$

where $G_1^{(D)}(x)$ and $G_2^{(D)}(x)$ are computable functions.

Moreover, $\text{Var}(X_n^{(D)})$ is positive for all $n$ large enough and we have

$$
\frac{X_n^{(D)} - \mathbb{E}(X_n^{(D)})}{\sqrt{\text{Var}(X_n^{(D)})}} \xrightarrow{d} N(0, 1).
$$
Remark 6. As stated in the above theorem, both functions $G_1^{(D)}(x)$ and $G_2^{(D)}(x)$ can be made explicit. More precisely, they both admit a representation as a double-integral which can be further simplified. We demonstrate this for the first function which is given by

$$G_1^{(D)}(\omega) = \int_0^\infty \frac{s^{\omega-1}}{Q(-2s)} \left( \int_0^\infty \tilde{f}(z)e^{-sz}dz \right) ds,$$

where $\tilde{f}(z) = (z^2/4 - 1)e^{-z} - ze^{-z/2} + 1$ and $Q(s) = \prod_{l\geq 1} (1 + s 2^{-l})$.

The inner integral is easy to evaluate

$$\int_0^\infty \tilde{f}(z)e^{-sz}dz = \frac{1}{2(s + 1)^3} - \frac{1}{s + 1} - \frac{1}{(s + 1/2)^2} + \frac{1}{s}.$$

Evaluating the outer integral is an exercise in $q$-analysis and Mellin transform and one obtains

$$G_1^{(D)}(\omega) = \kappa(-\omega)\Gamma(\omega)(1 - \omega) + \frac{Q(2^{\omega-1})}{Q(1)}\Gamma(-\omega)\Gamma(\omega + 1),$$

where at $\omega = 2$ this function is understood to be its limit and

$$\kappa(\omega) = \frac{8 \cdot 2^l - 32 \cdot 2^l + 46 \cdot 2^l - 32 \cdot 2^l + 9}{2^{1-l} \omega(2 \cdot 2^l - 1)^2(2^l - 1)^3} - \frac{2^{\omega+l+3} (\omega(2^{l+1} - 1) + 2^{l+1} - 2)}{(2 \cdot 2^l - 1)^2} + \frac{2^l (2^l \omega^2 + 3\omega - 2) - 2^{l+1}(\omega^2 + 4\omega - 2) + \omega^2 + 5\omega + 2}{4(2^l - 1)^3}.$$

This, e.g, yields

$$\frac{G_1^{(D)}(2)}{\log 2} \approx 0.30707981393605921828 \cdots$$

which coincides with the value given in [3].


