Adv. Appl. Prob. **49,** 850–876 (2017)
*doi:*10.1017/apr.2017.24

ON A RANDOM SEARCH TREE: ASYMPTOTIC ENUMERATION OF VERTICES BY DISTANCE FROM LEAVES

MIKLÓS BÓNA,* *University of Florida*
BORIS PITTEL,** *The Ohio State University*

Abstract

A random binary search tree grown from the uniformly random permutation of $[n]$ is studied. We analyze the exact and asymptotic counts of vertices by rank, the distance from the set of leaves. The asymptotic fraction c_k of vertices of a fixed rank $k \geq 0$ is shown to decay exponentially with k. We prove that the ranks of the uniformly random, fixed size sample of vertices are asymptotically independent, each having the distribution $\{c_k\}$. Notoriously hard to compute, the exact fractions c_k have been determined for $k \leq 3$ only. We present a shortcut enabling us to compute c_4 and c_5 as well; both are ratios of enormous integers, the denominator of c_5 being 274 digits long. Prompted by the data, we prove that, in sharp contrast, the largest prime divisor of the denominator of c_k is at most $2^{k+1} + 1$. We conjecture that, in fact, the prime divisors of every denominator for $k > 1$ form a single interval, from 2 to the largest prime not exceeding $2^{k+1} + 1$.

Keywords: Search tree; root; leaves; rank; enumeration; asymptotic; distribution; numerical data

2010 Mathematics Subject Classification: Primary 05A05
Secondary 05A15; 05A16; 05C05; 06B05;
05C80; 05D40; 60C05

1. Introduction

1.1. Background and definitions

Various parameters of the many models of random rooted trees are fairly well understood *if they relate to a near-root part of the tree or to a global tree structure*. The first group includes, for instance, the numbers of vertices at given distances from the root, the immediate progeny sizes for vertices near the top, and so on. See [8] for a comprehensive treatment of these results. The tree height and width are parameters of a global nature; see, for example, [3], [11]–[13], [17], and [18]. *Profiles* of random trees have been studied in [6] and [16]. In recent years there has been a growing interest in analysis of the random tree fringe, i.e. the tree part close to the leaves; see, [1], [2], [5], [9], [10], [14], and [15]. Diversity of models and techniques notwithstanding, a salient feature of these studies is usage of the inherently recursive nature of the random trees in question. Deletion of the root of a tree produces a forest of rooted subtrees that are conditionally independent, each being distributed as the random tree for the properly chosen tree size.

Received 21 March 2016; revision received 6 December 2016.

[*] Postal address: Department of Mathematics, University of Florida, 358 Little Hall, Gainesville, FL 32611-8105, USA. Email address: bona@ufl.edu

[**] Postal address: Department of Mathematics, The Ohio State University, 231 West 18th Avenue, Columbus, OH 43210-1175, USA. Email address: bgp@math.ohio-state.edu

Not surprisingly, the technical details of fringe analysis become quite complex as soon as the focus shifts to layers of vertices further away from the leaves. So while there are explicit results on the (limiting) fraction of vertices at a fixed, small distance from the leaves, an asymptotic behavior of this fraction, as a function of the distance, remains an open problem. In this paper we will solve this problem for the random *decreasing binary trees*, known also as *binary search trees*. We hope to study other random trees in a subsequent paper.

A decreasing binary tree on vertex set $[n] = \{1, 2, \dots, n\}$ is a binary plane tree in which every vertex has a smaller label than its parent. Note that this means that the root must have label n. Also note that every vertex has at most two children, and that every child v is either a left child or a right child of its parent, even if v is the only child of its parent.

Decreasing binary trees on vertex set $[n]$ are in bijection with permutations of $[n]$. In order to see this, let $p = p_1 p_2 \cdots p_n$ be a permutation. The decreasing binary tree of p, which we denote by $T(p)$, is defined as follows. The root of $T(p)$ is a vertex labelled n, the largest entry of p. If a is the largest entry of p on the left of n, and b is the largest entry of p on the right of n, then the root will have two children, the left one will be labelled a, and the right one labelled b. If n is the first (respectively, last) entry of p then the root will have only one child, and that is a left (respectively, right) child, and it will necessarily be labelled $n - 1$, as $n - 1$ must be the largest of all remaining elements. Define the rest of $T(p)$ recursively, by taking $T(p')$ and $T(p'')$, where p' and p'' are the substrings of p on the two sides of n, and affixing them to a and b.

1.2. Recent results

For the rest of this paper, whenever we say *tree*, we will mean a decreasing binary tree.

If v is a vertex of a tree T then let the *rank* of v be the number of edges in the shortest path from v to a leaf of T that is a descendant of v. So leaves are of rank 0, neighbors of leaves are of rank 1, and so on. Motivated by a series of recent papers [7], [14] concerning the neighbors of leaves, Bóna [2] proved that, for any $k \geq 0$, the probability that a randomly selected vertex of a randomly selected tree is of rank k converges to a rational number c_k as n goes to ∞. He also computed that $c_0 = \frac{1}{3}$, $c_1 = \frac{3}{10}$, $c_2 = \frac{1721}{8100}$, and $c_3 \approx 0.105$. It is worth mentioning that following this, Devroye and Janson [5] computed the same four values of c_k with a completely different method based on the ideas and techniques of branching processes. While the existence of c_k for all k was established in both [2] and [5] for all $k \geq 0$, the exact values of c_k for $k = 4, 5$, say, appeared to be out of reach.

The existing studies left wide open a series of very basic questions. Obviously, $\sum_k c_k \leq 1$, but is $\{c_k\}$ a probability distribution, i.e. $\sum_k c_k = 1$, and if yes, is $\{c_k\}$ the limiting distribution of the rank of a uniformly random vertex of the tree? What about the limiting *joint* distribution of the ranks of the random, fixed size, sample of the tree vertices? At exactly what speed does c_k approach 0? Is there a chance that a better understanding of the random tree structure can be used to compute, exactly, the constants c_k for some $k > 3$?

1.3. Main results

In this paper we are able to answer these questions. As in [2], our proofs continue to use the nonlinear, quadratic recurrences for the generating functions of counts of vertices with a given rank. To extract the required estimates from these recurrences, we use auxiliary enumerative schemes that lead to the linear recurrences being eminently amenable to asymptotic analysis. We feel confident that, properly modified, this approach can be applied to other models of random trees.

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

Theorem 1.1. (i) *The equality* $\sum_{k\geq 0} c_k = 1$ *holds, and so* $\{c_k\}$ *is the probability distribution of a random variable* R.

(ii) *Let* R_n *be the rank of the uniformly random vertex of the tree. Then for every* $0 < \rho < \frac{3}{2}$, *we have* $\lim_{n\to\infty} \mathbb{E}[\rho^{R_n}] = \mathbb{E}[\rho^R] < \infty$. *Consequently,* $R_n \to R$ *in distribution, and with all its moments, and* $c_k = O(q^k)$ *for every* $0 < q < \frac{2}{3}$.

(iii) *Let* $R_n^{(1)}, \dots, R_n^{(t)}$ *be the ranks of the uniformly random* t*-tuple of vertices of the tree. Then* $(R_n^{(1)}, \dots, R_n^{(t)})$ *converges in distribution to* $(R^{(1)}, \dots, R^{(t)})$ *with the components* $R^{(j)}$ *being independent copies of* R.

(iv) *Consequently, for each* k, *the fraction of vertices of rank* k *converges in probability to* c_k.

Part (ii) is consistent, broadly, with the conjecture in [2] stating that the sequence $\{c_k\}$ is log-concave. Focusing exclusively on this sequence we show that the decay of c_k is exactly exponentially fast.

To state the result concisely, introduce the function $g(\alpha) = \alpha + \alpha \log(2/\alpha) - 1$. The equation $g(\alpha) = 0$ has two positive roots. Let α_0 denote the smaller root; $\alpha_0 \approx 0.373$.

Theorem 1.2. *There exists* $\gamma > 0$ *such that, for all* $k \geq 1$,

$$\gamma \mathrm{e}^{-k/\alpha_0} \leq 1 - \sum_{j=0}^{k-1} c_j \leq \frac{6k+7}{3}\left(\frac{1}{3}\right)^k.$$

Note that if $\lim k^{-1} \log(1/c_k)$ exists, and we conjecture it does, then this limit is in $[\log 3, 1/\alpha_0]$.

In the course of proving these theorems, we stumbled on a shortcut in the computation of c_k described in [2]. It enabled us to obtain the precise values of c_4 and c_5, thus going beyond $c_0, \dots, c_3$ already found in [2] and [5].

Our numerical results and Theorem 1.1(iv) taken together show that, in a random binary search tree, with high probability, about 99.875 percent of all vertices are of rank 5 or less. When written in their simplest form, the numerators and denominators of the rational numbers c_k grow very fast. For instance, the denominator of c_5, which we denote by $\mathrm{denom}(c_5)$, has 274 digits. Despite its enormity, the largest prime divisor of $\mathrm{denom}(c_5)$ is 61. We conjectured and proved that this remarkable pattern holds for all k: the largest prime divisor of $\mathrm{denom}(c_k)$ is at most $2^{k+1} + 1$. So the 274-digit denominator of c_5 has no prime divisor larger than 65, i.e. larger than 61, which is indeed its prime divisor!

On the basis of our data, we conjecture that, for $k \geq 2$, the set of prime divisors of $\mathrm{denom}(c_k)$ is an *uninterrupted* interval of primes from 2 to the largest prime divisor, thus (by the prime number theorem) having length $\approx 2^{k+1}/(k \log 2)$ for large k. If true, this probably means that $\mathrm{denom}(c_k)$ is a product of certain factorials, hinting at some auxiliary enumerative scheme taking over at $n = \infty$.

That same data makes us believe that the numerator and the denominator of c_k are comparable in order of magnitude, but the numerator has very few prime factors, with the smallest one rapidly growing as k increases.

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

2. Convergence of the random rank R_n

We start by introducing $\mathbb{E}_{n,k}$, the expected number of vertices of rank k. Our focus is on the existence and the values of the limits

$$c_k = \lim_{n\to\infty} \frac{\mathbb{E}_{n,k}}{n}, \qquad k \geq 0.$$

Equivalently, c_k is the limiting probability that R_n, the rank of a uniformly random vertex of the (uniformly) random tree is k.

The data on c_k that we mentioned in Section 1.2 makes plausible a conjecture that $\{c_k\}$ is actually a probability distribution, so that there exists a random variable R such that $\mathbb{P}(R = k) = c_k$ and $R_n \to R$ in distribution. Our first theorem confirms this conjecture with room to spare, demonstrating that the moment generating function of R_n converges to that of R for any argument below $\frac{3}{2}$.

Theorem 2.1. *For every $\rho < \frac{3}{2}$, we have $\limsup \mathbb{E}[\rho^{R_n}] < \infty$. Consequently, $\{c_k\}$ is a probability distribution of a random variable R and $\lim \mathbb{E}[\rho^{R_n}] = \mathbb{E}[\rho^{R}]$.*

Proof. Let $p_{n,k}$ be the probability that the root is of rank k. Setting $\mathbb{E}_{0,k} \equiv 0$, we have

$$\mathbb{E}_{n,k} = p_{n,k} + \frac{1}{n}\sum_{j=0}^{n-1}(\mathbb{E}_{j,k} + \mathbb{E}_{n-1-j,k}), \qquad n \geq 1. \tag{2.1}$$

For $n = 1$, this equation holds trivially, since $\mathbb{E}_{1,k} = p_{1,k} = \mathbf{1}_{\{k=1\}}$. For $n > 1$, the equation holds because (2.1) just adds the expected value of the indicator of the event 'root is of rank k' to the expected total count of the nonroot vertices of rank k, the latter being first computed for trees in which the left subtree of the root is of size j. The existence of $c_k := \lim \mathbb{E}_{n,k}/n$, rational or not, will follow immediately from the next lemma.

Lemma 2.1. *Let $\{x_n\}_{n\geq 0}$, $\{y_n\}_{n\geq 1}$, and $\varepsilon \in (0, 1)$ be such that $x_0 = 0$, $y_n = O(n^{1-\varepsilon})$, and*

$$x_n = y_n + \frac{1}{n}\sum_{j=0}^{n-1}(x_j + x_{n-1-j}), \qquad n \geq 1.$$

Then there exists a finite $\lim_{n\to\infty} x_n/n$.

Proof. First of all, (2.1) is equivalent to

$$x_n = y_n + \frac{2}{n}\sum_{j=0}^{n-1}x_j, \qquad n \geq 1.$$

Standard manipulation shows that

$$nx_n - (n+1)x_{n-1} = ny_n - (n-1)y_{n-1}, \qquad n \geq 1,$$

or

$$\begin{aligned}\frac{x_n}{n+1} - \frac{x_{n-1}}{n} &= \frac{y_n}{n+1} - \frac{y_{n-1}}{n}\frac{n-1}{n+1} \\ &= \frac{ny_n - (n-1)y_{n-1}}{n(n+1)} \\ &= \frac{y_n}{n+1} - \frac{y_{n-1}}{n} + O(n^{-1-\varepsilon}), \qquad n \geq 1. \end{aligned} \tag{2.2}$$

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

Telescoping, we obtain, for $1 < m < n$,

$$\frac{x_n}{n+1} - \frac{x_m}{m+1} = \frac{y_n}{n+1} - \frac{y_m}{m+1} + O(m^{-\varepsilon}) = O(m^{-\varepsilon}).$$

Thus, $\{x_n/(n+1)\}$ is a fundamental Cauchy sequence, whence there exists a finite $\lim_{n\to\infty} x_n/(n+1)$, likewise $\lim_{n\to\infty} x_n/n$. □

Note that dropping the middle expression in (2.2) and adding the resulting equations, we obtain

$$x_n = (n+1)\sum_{j=1}^{n} \frac{jy_j - (j-1)y_{j-1}}{j(j+1)}, \tag{2.3}$$

which will come in handy later.

Let us proceed with the proof of Theorem 2.1. Since $p_{n,k} = O(1)$, the conditions of Lemma 2.1 obviously hold for $x_n = \mathbb{E}_{n,k}$ and $y_n = p_{n,k}$ with $\varepsilon \in (0, 1]$. Consequently, for each $k \ge 0$, there exists a finite limit $c_k := \lim \mathbb{E}_{n,k}/n$. Further, we have

$$\sum_k \frac{\mathbb{E}_{n,k}}{n} = 1 \quad\Longrightarrow\quad \sum_k c_k \le 1.$$

Next, given $\rho > 1$, introduce

$$\mathcal{H}_n(\rho) = \sum_{k\le n-1} \rho^k \mathbb{E}_{n,k},$$

which is the expected value of $\sum_{v\in[n]} \rho^{R(v)}$, where $R(v)$ denotes the rank of a generic vertex v. Then, analogously to (2.1),

$$\mathcal{H}_n(\rho) = \mathscr{h}_n(\rho) + \frac{\mathscr{L}}{n}\sum_{j=\mathscr{A}}^{n-\mathscr{L}} (\mathcal{H}_j(\rho) + \mathcal{H}_{n-\mathscr{L}-j}(\rho)), \qquad n > \mathscr{L}, \tag{2.4}$$

where $h_n(\rho) = \mathbb{E}[\rho^{R(\text{root})}]$. How large are $h_n(\rho)$ and $\mathcal{H}_n(\rho)$?

Let $X_{n,j}$ denote the random number of leaves at (edge) distance j from the root; $L_n = \sum_j X_{n,j}$ is the total number of leaves. Then

$$\rho^{R(\text{root})} \le \frac{\sum_j \rho^j X_{n,j}}{L_n} \quad\Longrightarrow\quad h_n(\rho) \le \mathbb{E}\left[\frac{\sum_j \rho^j X_{n,j}}{L_n}\right]. \tag{2.5}$$

We will show that L_n is of order n so it is likely that $h_n(\rho)$ is at most of order $n^{-1}\sum_j \rho^j \mathbb{E}[X_{n,j}]$. So let us bound $\sum_j \rho^j \mathbb{E}[X_{n,j}]$. To this end, attach to the random tree 'external' vertices, so that every vertex of the tree itself has exactly two descendants; thus, every leaf ℓ gets two external descendants, and every nonleaf vertex of the tree with one (left/right) descendant gets an additional external (right/left) descendant. Let $\mathcal{X}_{n,j}$ denote the total number of external nodes at distance j from the root. It was shown in [13] that

$$\mathcal{M}_j(x) := \sum_{n\ge\mathscr{L}} \mathbb{E}[\mathcal{X}_{n,j}]x^n = \frac{\mathscr{G}^j}{j!}\left(\log\frac{\mathscr{L}}{\mathscr{L}-x}\right)^j, \qquad j > \mathscr{A}.$$

Introduce $M_j(x) = \sum_{n\ge 0} x^n \mathbb{E}[X_{n,j}]$; so $M_0(x) = x$. Arguing as in [13], it can be shown that, for $j \ge 2$,

$$\frac{\mathrm{d}M_j(x)}{\mathrm{d}x} = \frac{2}{1-x} M_{j-1}(x).$$

Note that $[x^n]M_0(x) \le [x^n]\log(1/(1-x))$ for every $n \ge 0$. By induction on j, it follows that, for $j > 0$,

$$\mathbb{E}[X_{n,j}] = [x^n]M_j(x) \le \frac{2^j}{j!}[x^n]\biggl(\log\frac{1}{1-x}\biggr)^j = \mathbb{E}[\mathcal{X}_{n,\mathcal{j}}]. \tag{2.6}$$

Therefore, for every $r > 0$,

$$\begin{aligned}
\sum_{j\ge 0} r^j\mathbb{E}[X_{n,j}] &= [x^n]\sum_{j\ge 0} r^j M_j(x) \\
&\le [x^n]\sum_{j\ge 0} r^j \mathcal{M}_{\mathcal{j}}(x) \\
&= [x^n]\sum_{j\ge 0}\frac{(2r)^j}{j!}\biggl(\log\frac{1}{1-x}\biggr)^j \\
&= [x^n]\exp\biggl[2r\log\frac{1}{1-x}\biggr] \\
&= [x^n](1-x)^{-2r} \\
&= \binom{n+2r-1}{n} \\
&= \frac{\Gamma(n+2r)}{\Gamma(n+1)\Gamma(2r)} \\
&= O(n^{2r-1}),
\end{aligned}$$

the last equality following from the Stirling formula for the gamma function. Thus, for $r > 0$,

$$\sum_j r^j\mathbb{E}[X_{n,j}] = O(n^{2r-1}). \tag{2.7}$$

Consequently, for the numerator in the bound (2.5) of $h_n(\rho)$, we have

$$\mathbb{E}\biggl[\sum_j \rho^j X_{n,j}\biggr] = O(n^{2\rho-1}).$$

It remains to show that the denominator L_n in (2.5) is quite likely to be of order n, so that $h_n(\rho) = O(n^{2\rho-1}/n) = O(n^{2\rho-2})$. To be more specific, it was shown in [4] that $\mathbb{E}[L_n] = (n+1)/3$, so we should expect that $\mathbb{P}(L_n < an)$ is very small if $a < \frac{1}{3}$.

Lemma 2.2. *If* $x \in (0, 1]$ *and* $y \in (0, y(x))$*, where*

$$y(x) := (2\sqrt{1-x})^{-1}\log\frac{1+\sqrt{1-x}}{1-\sqrt{1-x}},$$

then, setting $L_0 = 0$*, we obtain*

$$\sum_{n\ge 0} y^n\mathbb{E}[x^{L_n}] = \sqrt{1-x}\,\frac{1+\mathrm{e}^{2y\sqrt{1-x}}((1-\sqrt{1-x})/(1+\sqrt{1-x}))}{1-\mathrm{e}^{2y\sqrt{1-x}}((1-\sqrt{1-x})/(1+\sqrt{1-x}))}. \tag{2.8}$$

Proof. Since, for $n > 1$,

$$\mathbb{E}[x^{L_n}] = \frac{1}{n}\sum_{k=0}^{n-1}\mathbb{E}[x^{L_k}]\mathbb{E}[x^{L_{n-1-k}}],$$

we obtain

$$\frac{\partial}{\partial y}\sum_{n\ge 0} y^n\mathbb{E}[x^{L_n}] = x+\sum_{n\ge 2} y^{n-1}\sum_{k=0}^{n-1}\mathbb{E}[x^{L_k}]\mathbb{E}[x^{L_{n-1-k}}] = \biggl(\sum_{n\ge 0} y^n\mathbb{E}[x^{L_n}]\biggr)^2 -(1-x). \quad (2.9)$$

Integrating and using $\sum_{n\ge 0} y^n\mathbb{E}[x^{L_n}]|_{y=0} = 1$, we obtain (2.8), provided that the denominator in (2.8) is positive, a condition equivalent to $y < y(x)$. □

Corollary 2.1. *Let $a < \frac{1}{3}$. For $\delta \in (0, 1)$, $b \in (a, \frac{1}{3})$, and large enough n, we have*

$$\mathbb{P}(L_n < an) \le \exp(-(b-a)n^{1-\delta}).$$

Proof. We start with a Chernoff-type bound

$$\mathbb{P}(L_n < an) \le x^{-an}y^{-n}\sum_{\nu\ge 0} y^{\nu}\mathbb{E}[x^{L_\nu}] \quad \text{for all } x < 1,\ y < y(x). \quad (2.10)$$

Choose $x = \exp(-n^{-\delta})$; then

$$y(x) = 1 + \frac{1}{3n^{\delta}} + O(n^{-2\delta}),$$

so we may choose $y = \exp(cn^{-\delta})$, $c \in (b, \frac{1}{3})$. From the right-hand side of (2.8), it follows that $\sum_n y^n\mathbb{E}[x^{L_n}] = O(n^{\delta})$. So, using also (2.8) and (2.10), we have

$$\mathbb{P}(L_n < an) = O[n^{\delta}\exp(an^{1-\delta} - cn^{1-\delta})] = O[\exp(-(b-a)n^{1-\delta})],$$

completing the proof of the corollary. □

Armed with (2.10), we return to (2.5). Let $D(\ell)$ denote the edge distance between the root and a generic leaf ℓ. By the Cauchy–Schwartz inequality,

$$\sum_j \rho^j X_{n,j} = \sum_{\ell}\rho^{D(\ell)} \le L_n^{1/2}\biggl(\sum_{\ell}\rho^{2D(\ell)}\biggr)^{1/2} \le n^{1/2}\biggl(\sum_j \rho^{2j}X_{n,j}\biggr)^{1/2}.$$

Therefore, applying the Cauchy–Schwartz inequality again and using (2.7),

$$\begin{aligned}
\mathbb{E}\biggl[\mathbf{1}_{\{L_n\le an\}}\sum_j \rho^j X_{n,j}\biggr] &\le n^{1/2}(\mathbb{E}[\mathbf{1}_{\{L_n\le an\}}])^{1/2}\biggl(\mathbb{E}\biggl[\sum_j \rho^{2j}X_{n,j}\biggr]\biggr)^{1/2}\\
&= n^{1/2}\mathbb{P}^{1/2}(L_n\le an)\biggl(\sum_j \rho^{2j}\mathbb{E}[X_{n,j}]\biggr)^{1/2}\\
&= O[n^{1/2}n^{(2\rho^2-1)/2}\mathbb{P}^{1/2}(L_n\le an)]\\
&= O[n^{\rho^2}\mathbb{P}^{1/2}(L_n\le an)].
\end{aligned}$$

Using the bound (2.7), with ρ^2 instead of ρ, and Corollary 2.1, we obtain

$$\mathbb{E}\biggl[\mathbf{1}_{\{L_n\le an\}}\sum_j \rho^j X_{n,j}\biggr] = O\biggl(n^{\rho^2}\exp\biggl(-\frac{(b-a)n^{1-\delta}}{4}\biggr)\biggr) = o(1).$$

Therefore, by (2.5) and (2.7),

$$h_n(\rho) \le \mathbb{E}\biggl[\mathbf{1}_{\{L_n< an\}}\sum_j \rho^j X_{n,j}\biggr] + \frac{1}{an}\sum_j \rho^j \mathbb{E}[X_{n,j}] = o(1) + O(n^{2\rho-2}). \tag{2.11}$$

Lemma 2.3. *For every fixed* $\rho < \frac{3}{2}$, *the limit* $\lim_{n\to\infty} n^{-1}\mathscr{H}_n(\rho)$ *exists and is finite. Consequently,* $\sum_{k\ge 0} c_k = 1$, $\sum_{k\ge 0}\rho^k c_k < \infty$, *and so* $c_k = o(\rho^{-k})$.

Proof. By (2.4) and (2.11), $x_n := \mathscr{H}_n(\rho)$ and $y_n := h_n(\rho)$ satisfy the condition of Lemma 2.1 with $\varepsilon \in (0, 3-2\rho]$. Hence, there exists a finite

$$\lim_{n\to\infty} n^{-1}\mathscr{H}_n(\rho) = \lim_{n\to\infty} n^{-\mathscr{E}}\mathbb{E}\biggl[\sum_{v\in[n]}\rho^{\mathcal{R}(v)}\biggr] = \lim_{n\to\infty} n^{-\mathscr{E}}\sum_{k\le n-\mathscr{E}}\rho^k\mathbb{E}_{n,k}.$$

Since $n^{-1}\sum_{k\le n-1}\mathbb{E}_{n,k} = 1$, and there exists $c_k = \lim_{n\to\infty} n^{-1}\mathbb{E}_{n,k}$, $(k \ge 0)$, we conclude that $\sum_k c_k = 1$, and

$$\lim_{n\to\infty} n^{-1}\sum_{0<k\le n-1}\rho^k\mathbb{E}_{n,k} = \sum_{k\ge 0}\rho^k c_k < \infty. \qquad \square$$

From Lemma 2.3, it follows that R_n, the rank $R(v)$ of the uniformly random vertex v, converges in distribution to R, $(\mathbb{P}(R=k) = c_k,\ k\ge 0)$ fast enough for $\mathbb{E}[\rho^{R_n}]$ to converge to $\mathbb{E}[\rho^R]$ if $\rho < \frac{3}{2}$. The proof of Theorem 2.1 is complete. □

Next we will show that the ranks of a finite ordered tuple of the random vertices are mutually independent in the limit $n\to\infty$.

Theorem 2.2. *Let* $t > 1$ *be fixed. For an ordered, fixed,* t*-tuple* $\boldsymbol{k} = (k_1,\dots,k_t)$, *let* $p_n(\boldsymbol{k})$ *denote the probability that the uniformly random* t*-tuple of vertices* $\boldsymbol{v} = (v_1,\dots,v_t)$ *have ranks* $R(v_1) = k_1,\dots,R(v_t) = k_t$. *Then* $\lim_{n\to\infty} p_n(\boldsymbol{k}) = \prod_{j=1}^t c_{k_j}$.

Proof. Let $\mathbb{E}_{n,\boldsymbol{k}}$ denote the expected number of t-tuples of vertices $v_1,\dots,v_t$ with ranks $k_1,\dots,k_t$, respectively; so $\mathbb{E}_{n,\boldsymbol{k}} = (n)_t\, p_n(\boldsymbol{k})$. So the claim is equivalent to

$$\mathbb{E}_{n,\boldsymbol{k}} = n^t\prod_{j=1}^t c_{k_j} + o(n^{t-1}).$$

For $t = 1$, the claim is obviously true. For $t \ge 2$, suppose that for $\tau < t$ and the tuples $(k_1,\dots,k_\tau)$, we have

$$\mathbb{E}_{n,(k_1,\dots,k_\tau)} = n^\tau\prod_{j=1}^\tau c_{k_j} + o(n^{\tau-1}).$$

Now we prove the claim for t-tuples in general.

First note that

$$\mathbb{E}_{n,\boldsymbol{k}} = \mathbb{E}'_{n,\boldsymbol{k}} + \mathbb{E}''_{n,\boldsymbol{k}};$$

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

here $\mathbb{E}'_{n,\boldsymbol{k}}$ is the contribution of the tuples $\boldsymbol{v}$ such that no v_i is a descendant of v_j, while $\mathbb{E}''_{n,\boldsymbol{k}}$ comes from the remaining tuples $\boldsymbol{v}$. Let us introduce the notation $(a)_b = a(a-1)\cdots(a-b+1)$. Obviously, and rather crudely,

$$\mathbb{E}''_{n,\boldsymbol{k}} \le (t)_2(n-2)_{t-2}\mathcal{E}_n,$$

where $\mathcal{E}_n$ is the expected number of pairs $\boldsymbol{v} = (v_1, v_2)$ such that v_2 is a descendant of v_1. Now

$$\mathcal{E}_n = n - 1 + \frac{2}{n}\sum_{j=0}^{n-1}\mathcal{E}_j.$$

Indeed, $n-1$ is the total number of the pairs (v_1, v_2) with $v_1 = n$, i.e. v_1 being at the root of the tree. As $\mathcal{E}_n = O(n\log n)$, we see that

$$\mathbb{E}''_{n,\boldsymbol{k}} = O(n^{t-1}\log n). \tag{2.12}$$

We emphasize that this simple bound holds independently of the claim we are proving. Now we turn to $\mathbb{E}'_{n,\boldsymbol{k}}$. Here none of the vertices from the tuple $\boldsymbol{v}$ can be at the root of the whole tree. So there are two distinct possibilities for the tuples $\boldsymbol{v} = (v_1, \dots, v_t)$:

- all v_j are contained in exactly one of two subtrees rooted at the two children of the root;
- there is a nonempty, proper subset $A \subset [t]$ such that the vertices v_j are in the left subtree if and only if $j \in [A]$.

Thus, denoting $\boldsymbol{k}_A = \{k_j,\ j \in A\}$, and $\boldsymbol{k}_{A^c} = \{k_j,\ j \in [t] \setminus A\}$,

$$\mathbb{E}'_{n,\boldsymbol{k}} = \frac{2}{n}\sum_{j=0}^{n-1} E'_{j,\boldsymbol{k}} + \frac{1}{n}\sum_{j=1}^{n-2}\sum_{\varnothing\ne A\subset[t]}\mathbb{E}'_{j,\boldsymbol{k}_A}\mathbb{E}'_{n-j-1,\boldsymbol{k}_{A^c}}. \tag{2.13}$$

Indeed, conditioned on the size of left subtree, the two subtrees are uniformly random decreasing trees of sizes j and $n-1-j$, respectively, and the number of tuples with ranks $\boldsymbol{k}_A$ in this tree and the number of tuples with ranks $\boldsymbol{k}_{A^c}$ in the right subtree are independent. Hence, the product of the expectations in the second sum. The first sum accounts for tuples $\boldsymbol{v}$ that lie entirely in just one of the two subtrees.

By (2.12) and the inductive assumption,

$$\begin{aligned}\mathbb{E}'_{j,\boldsymbol{k}_A} &= \mathbb{E}_{j,\boldsymbol{k}_A} - O(j^{|A|-1}\log j) = j^{|A|}\prod_{i\in A}c_{k_i} + o(j^{|A|}),\\ \mathbb{E}'_{n-1-j,\boldsymbol{k}_{A^c}} &= \mathbb{E}_{n-1-j,\boldsymbol{k}_{A^c}} - O((n-j)^{|A^c|-1}\log(n-j))\\ &= (n-1-j)^{|A^c|}\prod_{i\in A^c}c_{k_i} + o((n-j)^{|A^c|}).\end{aligned}$$

It follows easily that, for each A in question,

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n-2}\mathbb{E}'_{j,\boldsymbol{k}_A}\mathbb{E}'_{n-j-1,\boldsymbol{k}_{A^c}} &= \frac{1}{n}\prod_{i=1}^{t}c_{k_i}\sum_{j=1}^{n-2} j^{|A|}(n-1-j)^{|A^c|} + o(n^t)\\ &= n^t\prod_{i=1}^{t}c_{k_i}\int_0^1 x^{|A|}(1-x)^{|A^c|}\,\mathrm{d}x + o(n^t).\end{aligned}$$

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

So, adding these expressions for all subsets A in question, the second sum on the right-hand side of (2.13) is equal to

$$n^t \prod_{i=1}^{t} c_{k_i} \int_0^1 \sum_{a=1}^{t-1} \binom{t}{a} x^a (1-x)^{t-a}\, \mathrm{d}x + o(n^t) = \frac{t-1}{t+1} n^t \prod_{i=1}^{t} c_{k_i} + o(n^t).$$

Therefore, for every $\varepsilon > 0$, there exists $B_1 = B_1(\varepsilon) > 0$ such that

$$\frac{1}{n} \sum_{j=1}^{n-2} \sum_{\varnothing \neq A \subset [t]} \mathbb{E}'_{j,\boldsymbol{k}_A} \mathbb{E}'_{n-j-1,\boldsymbol{k}_{A^c}} \le b_n^+ := \frac{t-1}{t+1} n^t \prod_{i=1}^{t} c_{k_i} + \varepsilon n^t + B_1. \tag{2.14}$$

This implies that $\mathbb{E}'_{n,\boldsymbol{k}} \le \mathcal{E}^+_{n,\boldsymbol{k}}$, where

$$\mathcal{E}^+_{n,\boldsymbol{k}} = b_n^+ + \frac{2}{n} \sum_{j=0}^{n-1} \mathcal{E}^+_{j,\boldsymbol{k}}, \qquad \mathcal{E}^+_{j,\boldsymbol{k}} = 0, \qquad j < t.$$

So, using (2.3),

$$\mathcal{E}^+_{n,\boldsymbol{k}} = (n+1) \sum_{j=2}^{n} \frac{j b_j^+ - (j-1) b_{j-1}^+}{j(j+1)};$$

here, by (2.14),

$$\begin{aligned} \frac{j b_j^+ - (j-1) b_{j-!}^+}{j(j+1)} &= \frac{((t-1)/(t+1) \prod_i c_{k_i} + \varepsilon)(j^{t+1} - (j-1)^{t+1}) + B_1}{j(j+1)} \\ &\le \biggl(\frac{t-1}{t+1} \prod_i c_{k_i} + \varepsilon\biggr)(t+1) j^{t-2} + B_1 j^{-2}. \end{aligned}$$

Therefore, summing over $j \in [2, n]$,

$$\mathbb{E}'_{n,\boldsymbol{k}} \le \mathcal{E}^+_{n,\boldsymbol{k}} \le (n^t + n^{t-1}) \biggl(\prod_i c_{k_i} + \varepsilon \frac{t+1}{t-1}\biggr) + (n+1) B_1 \frac{\pi^2}{6}.$$

Similarly, for every $\varepsilon > 0$, there exists $B_2 = B_2(\varepsilon) > 0$ such that

$$\mathbb{E}'_{n,\boldsymbol{k}} \ge (n^t + n^{t-1}) \biggl(\prod_i c_{k_i} - \varepsilon \frac{t+1}{t-1}\biggr) - O(n B_2).$$

Thus,

$$\mathbb{E}'_{n,\boldsymbol{k}} = n^t \prod_{i=1}^{t} c_{k_i} + o(n^t).$$

Combining this estimate with (2.12), we complete the proof of the induction step. □

Let $V_{n,k}$ be the total number of vertices of rank k in all binary search trees on n vertices. So $V_{n,0} = L_n$ is the total number of leaves.

Corollary 2.2. *We have $V_{n,k}/n \to c_k$ in probability. That is, for every $\varepsilon > 0$, we have $\mathbb{P}(|V_{n,k}/n - c_k| > \varepsilon) = o(1)$ as $n \to \infty$.*

Proof. We know that $\mathbb{E}[V_{n,k}]/n = \mathbb{E}_{n,k}/n \to c_k$, and we also know that $\mathbb{E}[V_{n,k}(V_{n,k} - 1)/n(n-1)] \to c_k^2$. It remains to apply Chebyshev's inequality. □

Note that the result $\mathbb{P}(|V_{n,k}/n - c_k| > \varepsilon) = o(1)$ in Corollary 2.2 would not be enough for us. Recall though that, for $L_n := V_{n,0}$, we were able to show (Corollary 2.1) that $V_{n,0} < (c_0 - \varepsilon)n$ with probability at most $\exp(-\varepsilon n^{1-\delta})$, and that is smaller than n^{-K} for all $K > 0$. We conjecture that the analogous property holds for all $V_{n,k}$. A weaker claim, analogously proved, will suffice for our needs in Section 3.

Lemma 2.4. *There exists an absolute constant* $\beta > 0$*, such that, for* $\delta < 1$ *and* $n \geq n(\delta)$*,*

$$\mathbb{P}\biggl(V_{n,k} < \frac{\beta n}{3}\biggr) \leq \exp\biggl(-\frac{\beta n^{1-\delta}}{2}\biggr).$$

Proof. We split the proof into two parts. (i) Clearly, $V_{n,k} \geq \mathcal{V}_{n,\mathcal{k}}$, which is the total number of vertex-to-leaf paths of length k such that every nonleaf vertex of the path has only one child. Introduce

$$F(x,y) = \sum_{n\geq 0} y^n \mathbb{E}[x^{\mathcal{V}_{n,\mathcal{k}}}], \qquad \mathcal{V}_{\mathcal{0},\mathcal{k}} := \mathcal{0}.$$

Obviously, $\mathcal{V}_{n,\mathcal{k}} \leq n$; so for $y < 1$ the series converges if $xy < 1$. In particular, $F(x,\frac{1}{2})$ is analytic for $|x| < 2$. As $F(1,\frac{1}{2}) = 2$, we have, for $x \to 1$,

$$F\bigl(x,\tfrac{1}{2}\bigr) = 2 + \alpha(x-1) + O((x-1)^2), \qquad \alpha = F'_x\bigl(1,\tfrac{1}{2}\bigr) = \sum_{n\geq 1} 2^{-n}\mathbb{E}[\mathcal{V}_{n,\mathcal{k}}] > \mathcal{0}. \tag{2.15}$$

Now $\mathcal{V}_{n,\mathcal{k}} = \mathcal{0}$ for $n \leq k$, $\mathcal{V}_{\mathcal{k}+\mathcal{1},\mathcal{k}} = \mathcal{1}$ (respectively, 0) with probability $2^k/(k+1)!$ (respectively, $1 - 2^k/(k+1)!$), and, for $n > k+1$,

$$\mathbb{E}[x^{\mathcal{V}_{n,\mathcal{k}}}] = \frac{1}{n}\sum_{j=0}^{n-1} \mathbb{E}[x^{\mathcal{V}_{\mathcal{j},\mathcal{k}}}]\mathbb{E}[x^{\mathcal{V}_{n-\mathcal{1}-\mathcal{j},\mathcal{k}}}].$$

It follows, after simple algebra, that, for $x < 1$ and $y > 0$ such that the series for $F(x,y)$ converges,

$$\frac{\partial}{\partial y}F(x,y) = F^2(x,y) - (1-x)y^k\frac{2^k}{k!},$$

blending with (2.9) for $k = 0$. Consequently, for $y \geq \frac{1}{2}$,

$$\frac{\partial}{\partial y}F(x,y) \leq F^2(x,y) - a(1-x), \qquad a = \frac{1}{k}!.$$

Introduce $G(x,y)$, $y \geq \frac{1}{2}$, the solution of

$$\frac{\partial}{\partial y}G(x,y) = G^2(x,y) - a(1-x), \qquad G\bigl(x,\tfrac{1}{2}\bigr) = F\bigl(x,\tfrac{1}{2}\bigr).$$

Integrating the last equation and using

$$G^2\bigl(x,\tfrac{1}{2}\bigr) - a(1-x) = F^2\bigl(x,\tfrac{1}{2}\bigr) - a(1-x) > 0,$$

it follows that $G(x,y)$ exists for $x < 1$, $y \in [\frac{1}{2}, y_1(x))$,

$$y_1(x) := \frac{1}{2} + (2\sqrt{a(1-x)})^{-1}\log\frac{F(x,1/2) + \sqrt{a(1-x)}}{F(x,1/2) - \sqrt{a(1-x)}}, \tag{2.16}$$

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

and it is given by

$$
\begin{aligned}
G(x,y) &= \sqrt{a(1-x)} \\
&\quad\times \frac{1+\mathrm{e}^{(2y-1)(a(1-x))^{1/2}}((F(x,1/2)-\sqrt{a(1-x)})/(F(x,1/2)+\sqrt{a(1-x)}))}{1-\mathrm{e}^{(2y-1)(a(1-x))^{1/2}}((F(x,1/2)-\sqrt{a(1-x)})/(F(x,1/2)+\sqrt{a(1-x)}))} \\
&= O\biggl(\frac{1}{y_1(x)-y}\biggr), \qquad y \uparrow y_1(x), \qquad (2.17)
\end{aligned}
$$

uniformly for $x < 1$. Consequently, $F(x,y)$ exists for $y < y_1(x)$, and $F(x,y) \le G(x,y)$ for $y \in [\frac{1}{2}, y_1(x))$. Using (2.15) and (2.16), we obtain, for $x \to 1$,

$$y_1(x) = 1 + (1-x)\beta + O((1-x)^2), \qquad \beta = \frac{\alpha}{4} + \frac{a}{24}. \qquad (2.18)$$

(ii) Armed with (2.17), (2.18), and $F(x,y) \le G(x,y)$, we choose $x = \mathrm{e}^{-n^{-\delta}}$ and $y = \mathrm{e}^{6\beta n^{-\delta}/7}$, which is strictly below $y_1(x)$ for large n, and apply the Chernoff-type bound, i.e.

$$
\begin{aligned}
\mathbb{P}\biggl(\mathcal{V}_{\boldsymbol{n},\boldsymbol{k}} < \frac{\boldsymbol{\beta n}}{\mathcal{G}}\biggr) &\le x^{-\beta n/3} y^{-n} F(x,y) \\
&\le x^{-\beta n/3} y^{-n} G(x,y) \\
&= O\biggl(n^{\delta} \exp\biggl(\frac{\beta n^{1-\delta}}{3} - \frac{6\beta n^{1-\delta}}{7}\biggr)\biggr) \\
&\le \exp\biggl(-\frac{\beta n^{1-\delta}}{2}\biggr).
\end{aligned}
$$

□

3. A closer look at the distribution $\{c_k\}$

In Theorem 2.1, we proved the existence of the finite $\lim_{n\to\infty} n^{-1}\sum_k \rho^k \mathbb{E}_{n,k}$ for $\rho < \frac{3}{2}$, which implied that $1 - \sum_{j=0}^{k-1} c_k = O(q^k)$ for every $q > \frac{2}{3}$. Focusing exclusively on the sequence $\{c_k\}$, we prove a considerably stronger bound.

Theorem 3.1. *The following inequality holds:*

$$1 - \sum_{j=0}^{k} c_j \le \frac{6k+7}{3}\biggl(\frac{1}{3}\biggr)^k.$$

Proof. We split the proof into four parts. (i) For $n \ge 1$, $k \ge 0$, let $a_{n,k}$ be the total number of vertices of rank k in all $n!$ permutations of $[n]$, and let $b_{n,k}$ be the total number of permutations for which the root of the tree is of rank k. So $a_{n,k}/n! = \mathbb{E}_{n,k}$, the expected number of rank-k vertices in the random tree, and $b_{n,k}/n!$ is the probability that its root is of rank k. Introduce $A_k(x) = \sum_{n>0} x^n a_{n,k}/n!$ and $B_k(x) = \sum_{n>0} x^n b_{n,k}/n!$; in particular, $B_0(x) = x$. From [2, Lemmas 3.1 and 3.2],

$$
\begin{aligned}
A_k'(x) &= \frac{2}{1-x}A_k(x) + B_k'(x), \qquad k \ge 0, \\
B_k'(x) &= 2B_{k-1}(x)\biggl(\frac{1}{1-x} - \sum_{j=0}^{k-2} B_j(x)\biggr) - B_{k-1}(x)^2, \qquad k > 0.
\end{aligned}
\qquad (3.1)
$$

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

Introduce $A_{\le k}(x) = \sum_{0\le j\le k} A_j(x)$ and $B_{\le k}(x) = \sum_{0\le j\le k} B_j(x)$. In particular, $A_{\le k}(x)$ is the generating function of $\{\sum_{j\le k}\mathbb{E}_{n,j}\}_{n\ge 0}$. It follows from (3.1) that

$$A'_{\le k}(x) = \frac{2}{1-x}A_{\le k}(x) + B'_{\le k}(x), \qquad k \ge 0, \tag{3.2}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\biggl(\frac{1}{1-x} - B_{\le k}(x)\biggr) = \biggl(\frac{1}{1-x} - B_{\le k-1}(x)\biggr)^2 - 1, \qquad k > 0. \tag{3.3}$$

Equation (3.3) can also be obtained directly via the conditional independence argument as follows. Let $p_{n,\le k}$ be the probability that the root rank is k at most, so $p_{n,>k} := 1 - p_{n,\le k}$ is the probability that the root rank strictly exceeds k. Clearly, $p_{n,\le k} = \sum_{j\le k} b_{n,j}/n!$, and, therefore, $B_{\le k}(x)$ is the generating function of $\{p_{n,\le k}\}_{n\ge 1}$. Then, for $n > 1$ and $k \ge 0$,

$$p_{n,>k} = \biggl(\frac{1}{n}\biggr)\sum_{j=0}^{n-1} p_{j,>k-1}\,p_{n-j-1,>k-1},$$

where $p_{0,>k-1} := 1$, since conditioned on the left subtree having size k, the left subtree and the right subtree are independent. Consequently, as $p_{1,>k} = 0$ for all $k \ge 0$,

$$\frac{\mathrm{d}}{\mathrm{d}x}\sum_{n\ge 1} p_{n,>k}x^n = \biggl(\sum_{n\ge 0} p_{n,>k-1}x^n\biggr)^2 - p_{0,>k-1}p_{0,>k-1} = \biggl(\sum_{n\ge 0} p_{n,>k-1}x^n\biggr)^2 - 1.$$

Here

$$\sum_{n\ge 1} p_{n,>k}x^n = \sum_{n\ge 1}(1 - p_{n,\le k})x^n = \frac{x}{1-x} - B_{\le k}(x) = \frac{1}{1-x} - 1 - B_{\le k}(x),$$

and

$$\sum_{n\ge 0} p_{n,>k-1}x^n = 1 + \sum_{n\ge 1} p_{n,>k-1}x^n = 1 + \frac{x}{1-x} - B_{\le k-1}(x) = \frac{1}{1-x} - B_{\le k-1}(x)$$

with $B_{\le -1}(x) := 0$.

So

$$\frac{\mathrm{d}}{\mathrm{d}x}\biggl(\frac{1}{1-x} - B_{\le k}(x)\biggr) = \biggl(\frac{1}{1-x} - B_{\le k-1}(x)\biggr)^2 - 1.$$

(ii) From (3.2), it follows that, for $k > 0$,

$$\begin{aligned} A_{\le k}(x) &= \frac{1}{(1-x)^2}\int_0^x (1-y)^2 B'_{\le k}(y)\,\mathrm{d}y \\ &= \frac{1}{(1-x)^2}\biggl[(1-x)^2 B_{\le k}(x) + 2\int_0^x (1-y)B_{\le k}(y)\,\mathrm{d}y\biggr]. \end{aligned}$$

So for $x \uparrow 1$, we have

$$A_{\le k}(x) \sim \frac{2}{(1-x)^2}\int_0^1 (1-y)B_{\le k}(y)\,\mathrm{d}y.$$

Since $A_{\le k}(x)$ is the generating function of $\{\sum_{j\le k}\mathbb{E}_{n,j}\}_{n>0}$ and $n^{-1}\sum_{j\le k}\mathbb{E}_{n,j}\to\sum_{j=0}^{k}c_j$, it follows by the Tauberian theorem that

$$\sum_{j=0}^{k}c_j=2\int_0^1(1-y)B_{\le k}(y)\,\mathrm{d}y. \tag{3.4}$$

Obviously,

$$B_{\le k}(x)=\frac{x}{1-x}-B_{>k}(x),$$

where $B_{>k}(x)$ is the generating function of $\{b_{n,>k}/n!\}$, $b_{n,>k}$ being the number of permutations such that the root rank (strictly) exceeds k. Consequently,

$$1-\sum_{j=0}^{k}c_j=2\int_0^1(1-y)B_{>k}(y)\,\mathrm{d}y. \tag{3.5}$$

Thus, to bound $1-\sum_{j=0}^{k}c_j$ from above we need to bound $B_{>k}(x)$ from above. Clearly, $b_{n,>k}$ is bounded above by the number of permutations for which there exists a root-to-leaf path of (edge) length exceeding k. The success of this approach depends on how efficient our search would be for finding a path that has a good chance to be comparable in length to the shortest path.

(iii) We introduce a randomized greedy algorithm with a plausibly good chance to find such a competitive path. If there are two nonempty subtrees at the root of the tree, we *delete* a subtree with probability proportional to the number of vertices in it. We repeat the same procedure at the root of the remaining subtree and continue until the remaining subtree is a leaf of the whole tree. The resulting sequence of roots of the nested subtrees forms a root-to-leaf path in the whole tree.

For $n\ge 1$, $k\ge -1$, let $\pi_{n,>k}$ denote the probability that the length of this path exceeds k; obviously, $p_{n,>k}\le\pi_{n,>k}$. Further, $\pi_{n,>-1}=1$, and, for $n>1$, $k\ge 0$,

$$\pi_{n,>k}=\frac{2}{n}\pi_{n-1,>k-1}+\frac{1}{n}\sum_{j=1}^{n-2}\biggl[\frac{n-1-j}{n-1}\pi_{j,>k-1}+\frac{j}{n-1}\pi_{n-1-j,>k-1}\biggr],$$

or

$$(n)_2\pi_{n,>k}=2(n-1)\pi_{n,>k-1}+2\sum_{j=1}^{n-2}(n-1-j)\pi_{j,>k-1}. \tag{3.6}$$

Introduce $\mathbb{P}_{>k}(x)=\sum_{n>0}\pi_{n,>k}x^n$; in particular,

$$\mathbb{P}_{>-1}(x)=\sum_{n>0}x^n=\frac{x}{1-x}.$$

Obviously, $B_{>k}(x)\le\mathbb{P}_{>k}(x)$, and so (3.5) yields

$$1-\sum_{j=0}^{k}c_j\le 2\int_0^1(1-y)\mathbb{P}_{>k}(y)\,\mathrm{d}y,\qquad k\ge 0. \tag{3.7}$$

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

Since $\pi_{n,>k} = 0$ for $n \le k$, we have $\mathbb{P}^{(t)}_{>k}(0) = 0$ for $t \le k$. It follows from (3.6) that

$$\begin{aligned}\frac{\mathrm{d}^2\mathbb{P}_{>k}(x)}{\mathrm{d}x^2} &= \sum_{n\ge 2}(n)_2\pi_{n,>k}x^{n-2}\\ &= 2\sum_{n\ge 2}(n-1)\pi_{n-1,>k-1}x^{n-2} + 2\sum_{n\ge 2}x^{n-2}\sum_{j=1}^{n-2}(n-1-j)\pi_{j,>k-1}\\ &= 2\frac{\mathrm{d}}{\mathrm{d}x}\sum_{\nu\ge 1}\pi_{\nu,>k-1}x^{\nu} + 2\sum_{j\ge 1}\pi_{j,>k-1}x^j\sum_{n\ge j+2}(n-1-j)x^{n-2-j}\\ &= 2\frac{\mathrm{d}\mathbb{P}_{>k-1}}{\mathrm{d}x} + 2\biggl(\sum_{j\ge 1}\pi_{j,>k-1}x^j\biggr)\biggl(\sum_{\nu\ge 1}\nu x^{\nu-1}\biggr)\\ &= 2\frac{\mathrm{d}\mathbb{P}_{>k-1}}{\mathrm{d}x} + \frac{2}{(1-x)^2}\mathbb{P}_{>k-1}(x).\end{aligned}$$

Thus,

$$\frac{\mathrm{d}^2\mathbb{P}_{>k}(x)}{\mathrm{d}x^2} = 2\frac{\mathrm{d}\mathbb{P}_{>k-1}}{\mathrm{d}x} + \frac{2}{(1-x)^2}\mathbb{P}_{>k-1}(x) \tag{3.8}$$

with $\mathbb{P}^{(r)}_{>k}(0) = 0$ for $r \le k$. In light of (3.7) it seems necessary, as before, to integrate successively the differential equations (3.8) for $\mathbb{P}_{>k'}(x)$, $k' = 1, 2, \dots, k$, and then to evaluate the right-hand side of the bound (3.7). In fact, that is how we computed the bounds (3.7) for k up to 10; linearity of (3.8) was critical for success of this computation. The data showed, rather compellingly, that the bound decays faster than $(\frac{1}{2})^k$. In the absence of any tractable expression for $\mathbb{P}_{>k}(x)$ when k is large, the issue was to find a way to bound the integral in (3.7) without such an expression.

(iv) Linearity of (3.8) to the rescue again! Introduce

$$I_{k,t} = \int_0^1 (1-y)^t\mathbb{P}_{>k}(y)\,\mathrm{d}y, \qquad k \ge -1,\ t > 0;$$

so

$$1 - \sum_{j=0}^{k-1} c_k \le 2I_{k,1}, \qquad k > 0. \tag{3.9}$$

First note that, for $t > 0$,

$$I_{-1,t} = \int_0^1 (1-y)^t\mathbb{P}_{>-1}(y)\,\mathrm{d}y = \int_0^1 [-(1-y)^t + (1-y)^{t-1}]\,\mathrm{d}t = \frac{1}{t(t+1)}. \tag{3.10}$$

Let us show that, for $k \ge 0$, $t > 0$,

$$I_{k,t} = \frac{2}{(t+2)_2}[I_{k-1,t} + (t+2)I_{k-1,t+1}]. \tag{3.11}$$

Indeed, using $\mathbb{P}^{(r)}_{>k}(0) = 0$ for $r = 0, 1$ and (3.8),

$$\begin{aligned}I_{k,t} &= \int_0^1 (1-y)^t\mathbb{P}_{>k}(y)\,\mathrm{d}y\\ &= \frac{1}{(t+2)_2}\int_0^1 (1-y)^{t+2}\frac{\mathrm{d}^2\mathbb{P}_{>k}(y)}{\mathrm{d}y^2}\,\mathrm{d}y\end{aligned}$$

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

$$
\begin{aligned}
&= \frac{2}{(t+2)_2}\int_0^1 (1-y)^{t+2}\biggl[\frac{\mathrm{d}\mathbb{P}_{>k-1}(y)}{\mathrm{d}y} + \frac{\mathbb{P}_{>k-1}(y)}{(1-y)^2}\biggr]\mathrm{d}y \\
&= \frac{2}{(t+2)_2}\biggl[(t+2)\int_0^1 (1-y)^{t+1}\mathbb{P}_{>k-1}(y)\,\mathrm{d}y + \int_0^1 (1-y)^t\mathbb{P}_{>k-1}(y)\,\mathrm{d}y\biggr] \\
&= \frac{2}{(t+2)_2}[I_{k-1,t} + (t+2)I_{k-1,t+1}].
\end{aligned}
$$

In particular,

$$I_{k,1} = \tfrac{1}{3}[I_{k-1,1} + 3I_{k-1,2}] \ge \tfrac{1}{3}I_{k-1,1},$$

so that $I_{k,1} \ge \text{constant}\times(\frac{1}{3})^k$. Next We show that, in fact, $I_{k,1} \le \text{constant}\times(\frac{1}{3})^k$, i.e. $I_{k,1}$ is of order $(\frac{1}{3})^k$ exactly.

To this end, fix $\tau > 0$ and consider $I_{k,t}$ for $k \ge -1$ and $t \ge \tau$. Let us show that

$$I_{k,t} \le \frac{1}{(t+1)_2}\biggl(\frac{2}{\tau+2}\biggr)^{k+1}. \tag{3.12}$$

By (3.10), the bound holds for $k = -1$. Inductively, if it holds for some $k \ge 0$ then by (3.11),

$$
\begin{aligned}
I_{k+1,t} &\le \frac{2}{(t+2)_2}\biggl[\frac{1}{(t+1)_2}\biggl(\frac{2}{\tau+2}\biggr)^{k+1} + \frac{t+2}{(t+2)_2}\biggl(\frac{2}{\tau+2}\biggr)^{k+1}\biggr] \\
&= \frac{2}{(t+2)_3}\biggl(\frac{2}{\tau+2}\biggr)^{k+1} \\
&\le \frac{1}{(t+1)_2}\biggl(\frac{2}{\tau+2}\biggr)^{k+2}.
\end{aligned}
$$

So the bound (3.12) is proven. In particular, for $t \ge 4$,

$$I_{k,t} \le \frac{1}{(t+1)_2}\biggl(\frac{1}{3}\biggr)^{k+1} \quad\Longrightarrow\quad I_{k,4} \le 0.05\biggl(\frac{1}{3}\biggr)^{k+1}.$$

Using (3.11) for $t = 3$, we have

$$I_{k,3} = \tfrac{1}{10}I_{k-1,3} + \tfrac{1}{2}I_{k-1,4} \le 0.1I_{k-1,3} + 0.025\bigl(\tfrac{1}{3}\bigr)^k.$$

Iterating this recurrence inequality and using (3.10) for $I_{-1,3}$, we obtain

$$
\begin{aligned}
I_{k,3} &\le \frac{1}{3\cdot 4}\biggl(\frac{1}{10}\biggr)^{k+1} + 0.025\biggl(\frac{1}{3}\biggr)^k\sum_{j\ge 0}\biggl(\frac{3}{10}\biggr)^j \\
&= \biggl(\frac{1}{3}\biggr)^{k+1}\biggl(\frac{1}{12} + 0.025\biggl(\frac{30}{7}\biggr)\biggr) \\
&\le \frac{1}{5}\biggl(\frac{1}{3}\biggr)^{k+1}.
\end{aligned} \tag{3.13}
$$

Analogously, using (3.11) for $t = 2$ in conjunction with (3.13), we iterate the resulting recurrence inequality

$$I_{k,2} \le \tfrac{1}{6}I_{k-1,2} + \tfrac{2}{15}\bigl(\tfrac{1}{3}\bigr)^k.$$

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

Recalling (3.10) for $I_{-1,2}$, we obtain

$$I_{k,2} \leq \left(\tfrac{1}{3}\right)^{k+1}. \tag{3.14}$$

Finally, combining (3.11) for $t = 1$ and (3.14), we have

$$I_{k,1} \leq \tfrac{1}{3} I_{k-1,1} + \left(\tfrac{1}{3}\right)^{k}.$$

Using this recurrence, and (3.10) for $I_{-1,1}$, we arrive at

$$I_{k,1} \leq \frac{6k+7}{6}\left(\frac{1}{3}\right)^{k}. \tag{3.15}$$

The bounds (3.9) and (3.15) taken together imply that

$$1 - \sum_{j=0}^{k} c_j \leq \frac{6k+7}{3}\left(\frac{1}{3}\right)^{k}. \qquad \square$$

Next we prove a qualitatively matching lower bound for $1 - \sum_{j=0}^{k} c_j$. Introduce the function $g(\alpha) = \alpha + \alpha \log(2/\alpha) - 1$. The equation $g(\alpha) = 0$ has two positive roots. Let α_0 denote the smaller root; $\alpha_0 \approx 0.373$. It was proved in [13] that the likely length of the shortest path from the root of the random tree to a leaf is at least $(\alpha_0 - \varepsilon)\log n$ for every $\varepsilon > 0$.

Theorem 3.2. *There exists a positive constant γ such that, for all $k \geq 0$,*

$$1 - \sum_{j=0}^{k} c_j \geq \gamma \mathrm{e}^{-k/\alpha_0}.$$

Proof. We split the proof into two parts. (i) Given an integer m, consider the random tree on $[m]$. Let S_m denote the edge length of the shortest path from the root to a leaf. Then, for every $s \in [0, m-1]$,

$$\mathbb{P}(S_m \leq s) = \sum_{\mu \leq s} \mathbb{P}(S_m = \mu) \leq \sum_{\mu \leq s} \mathbb{E}[X_{m,\mu}],$$

where $X_{m,\mu}$ is the total number of leaves at distance μ from the root. By (4.1), proved independently in the next section,

$$\mathbb{E}[X_{m,\mu}] \leq \frac{2^{\mu}}{(\mu-1)!} \frac{(\log m + 1)^{\mu-1}}{m}.$$

So, given $\gamma > 0$, we have, for $\mu \leq \gamma \log m$,

$$\mathbb{E}[X_{m,\mu}] \leq \frac{\mu}{m(\log m + 1)} \frac{2^{\mu}(\log m + 1)^{\mu}}{\mu!} \leq \frac{\gamma \mathrm{e}^{\gamma}}{m}\left(\frac{2\log m}{\mu/\mathrm{e}}\right)^{\mu}.$$

As a function of μ, the right-hand side increases for $\mu \leq 2\log m$. Assuming that $\gamma \leq 2$, we obtain

$$\mathbb{P}(S_m \leq \gamma \log m) \leq \frac{\gamma^2 \mathrm{e}^{\gamma} \log m}{m}\left(\frac{2\mathrm{e}}{\gamma}\right)^{\gamma \log m} = \gamma^2 \mathrm{e}^{\gamma} (\log m) \mathrm{e}^{g(\gamma)\log m}.$$

Now $g(\alpha)$ is strictly increasing on $[0, \alpha_0]$, from $g(0) = -1$ to $g(\alpha_0) = 0$. So picking $\gamma = \alpha_0/2$ say, we obtain

$$\mathbb{P}\left(S_m \leq \left(\frac{\alpha_0}{2}\right)\log m\right) \leq \left(\frac{\alpha_0}{2}\right)^2 \mathrm{e}^{\alpha_0/2} m^{g(\alpha_0/2)} \log m = o(1). \tag{3.16}$$

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

For $\alpha := \mu/\log m \in (\alpha_0/2, \alpha_0)$, we have (see [13])

$$\mathbb{E}[X_{m,\mu}] = (1 + O(\varepsilon_m))K(\alpha)(\log m)^{-1/2}\mathrm{e}^{g(\alpha)\log m}, \qquad K(\alpha) := (\sqrt{2\pi\alpha}\,\Gamma(\alpha))^{-1}\mathrm{e}^{\alpha-1},$$

where $\lim_{m\to\infty}\varepsilon_m = 0$. By the convexity of $g(\alpha)$ on $[0, \alpha_0]$,

$$g(\alpha) \le g(\alpha_0) + (\alpha - \alpha_0)g'(\alpha_0) = (\alpha - \alpha_0)g'(\alpha_0),$$

where $g' := g'(\alpha_0) > 0$. Therefore, $\sum_{\alpha\in(\alpha_0/2,\alpha_0]}\mathbb{E}[X_{m,\mu}]$ is of order

$$(\log m)^{-3/2}\int_{\alpha_0 g'/2}^{\alpha_0 g'} \mathrm{e}^{-x}\,\mathrm{d}x = O((\log m)^{-3/2}).$$

Recalling (3.16), we conclude that

$$\mathbb{P}(S_m \le \alpha_0 \log m) = O((\log m)^{-3/2}). \tag{3.17}$$

(ii) Given $\ell \ge 0$, let $Y_{n,\ell}$ denote the total number of subtrees of size $m \ge \ell$. Then, for $n \ge \ell$,

$$\mathbb{E}[Y_{n,\ell}] = 1 + \frac{1}{n}\sum_{j=0}^{n-1}(\mathbb{E}[Y_{j,\ell}] + \mathbb{E}[Y_{n-1-j,\ell}])$$

with $\mathbb{E}[Y_{j,\ell}] = 0$ for $j < \ell$. The standard computation shows that

$$\frac{\mathbb{E}[Y_{n,\ell}]}{n+1} = \frac{2}{\ell+1} - \frac{1}{n+1} \quad\Longrightarrow\quad \frac{\mathbb{E}[Y_{n,\ell}]}{n} = \biggl(1 + \frac{1}{n}\biggr)\biggl(\frac{2}{\ell+1} - \frac{1}{n+1}\biggr). \tag{3.18}$$

Consider a generic subtree on $m \ge \ell$ vertices. Conditioned on its vertex set $\{p(i_1), \dots, p(i_m)\}$, $i_1 < \cdots < i_m$, this subtree has the same distribution as the tree on $[m]$ grown from a uniformly random permutation of $[m]$. So, denoting $S(p(i_1), \dots, p(i_m))$ the length of the shortest root-to-leaf path in this subtree by (3.17), we have, uniformly for $m \ge \ell$,

$$\mathbb{P}(S(p(i_1), \dots, p(i_m)) > \alpha_0 \log \ell \mid p(i_1), \dots, p(i_m)) = 1 - O((\log \ell)^{-3/2}).$$

Let $Z_{n,\ell}$ denote the total number of the subtrees of size $m \ge \ell$ such that the shortest root-to-leaf path has length exceeding $\alpha_0 \log \ell$; clearly,

$$\sum_{j\ge\alpha_0\log\ell} \mathbb{E}_{n,j} \ge \mathbb{E}[Z_{n,\ell}].$$

From this equation, it follows that

$$\mathbb{E}[Z_{n,\ell} \mid Y_{n,\ell}] = [1 - O((\log \ell)^{-3/2})]Y_{n,\ell}.$$

Combining this with (3.18), we obtain

$$\frac{\mathbb{E}[Z_{n,\ell}]}{n} = \frac{2}{\ell+1}[1 + O((\log \ell)^{-3/2} + n^{-1})].$$

Therefore,

$$\sum_{j\ge\alpha_0\log\ell} c_j = \lim_{n\to\infty} n^{-1}\sum_{j\ge\alpha_0\log\ell}\mathbb{E}_{n,j} \ge \liminf_{n\to\infty}\frac{\mathbb{E}[Z_{n,\ell}]}{n} = \frac{2}{\ell+1}[1 + O((\log \ell)^{-3/2})].$$

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

Pick $k > 0$ and set $\ell = \lceil \mathrm{e}^{k/\alpha_0} \rceil$. Then the above estimate implies that

$$\sum_{j>k} c_j \geq \frac{2}{\lceil \mathrm{e}^{k/\alpha_0} \rceil + 1}[1 + O(k^{-3/2})] \geq \frac{2}{3}\mathrm{e}^{-k/\alpha_0}[1 + O(k^{-3/2})]. \qquad \square$$

Remark 3.1. By Theorems 3.1 and 3.2, the radius of convergence of $\sum_k c_k x^k$ is in the interval $[3, \mathrm{e}^{1/0.373\cdots}]$. What is the exact value of the radius?

4. Variations

Besides $\mathbb{E}_{n,k}$, the expected counts of rank-k vertices, it is also natural to consider $F_{n,k}$ and $G_{n,k}$, the expected number of all pairs (v, u), where v is a vertex of rank k and u is a descendant leaf of v and the expected number of all pairs (v, u), where v is a vertex of rank k and u is a *closest* descendant leaf of v.

Let us show that, for each k, there exist finite limits $f_k = \lim_{n\to\infty} F_{n,k}/n$, and $g_k = \lim_{n\to\infty} G_{n,k}/n$. Consider $F_{n,k}$, for example. Introducing $f_{n,k}$, the expected product of the number of leaves of the random tree and the indicator of the event {root rank $= k$}, we have

$$F_{n,k} = f_{n,k} + \left(\frac{1}{n}\right)\sum_{j=0}^{n-1}(F_{j,k} + F_{n-1-j,k}), \qquad n > 1.$$

For $n > 0$, $f_{n,0} = 0$; for $k > 0$, using (2.6),

$$\begin{aligned}
f_{n,k} &\leq (n-1)\mathbb{P}(\text{root rank} = k) \\
&\leq n\mathbb{E}[X_{n,k}] \\
&\leq n2^k[x^n]\frac{1}{k!}\left(\log\frac{1}{1-x}\right)^k \\
&= 2^k\frac{n}{n!}[y^{k-1}](y+1)\cdots(y+n-1) \\
&= 2^k\frac{n(n-1)!}{n!}\left(\sum_{0<i_1<\cdots<i_{k-1}<n}\frac{1}{i_1\cdots i_{k-1}}\right) \\
&\leq \frac{2^k}{(k-1)!}\left(\sum_{1\leq i\leq n-1}\frac{1}{i}\right)^{k-1} \\
&\leq \frac{2^k}{(k-1)!}(\log n + 1)^{k-1} \\
&= O((\log n)^{k-1}).
\end{aligned} \tag{4.1}$$

So $x_n := F_{n,k}$ and $y_n := f_{n,k}$ meet the conditions of Lemma 2.1 with $\varepsilon \in (0, 1)$. Consequently, for each k, there exists a finite $f_k := \lim_{n\to\infty} F_{n,k}/n$.

To compute f_k and g_k, we need the recurrences similar to (3.2) and (3.3). Introduce $f_{n,>k} = \sum_{j>k} f_{n,j}$, and $\mathcal{A}_{\mathcal{k}}(x) = \sum_{n\geq 1} x^n F_{n,k}$, $\mathcal{B}_{\mathcal{k}}(x) = \sum_{n\geq 1} x^n f_{n,k}$, and $\mathcal{B}_{>\mathcal{k}}(x) = \sum_{n\geq 1} x^n f_{n,>k}$. Then $\mathcal{B}_{\mathcal{k}}(x) = \mathcal{B}_{>\mathcal{k}-\mathcal{1}}(x) - \mathcal{B}_{>\mathcal{k}}(x)$.

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

Lemma 4.1. *For all nonnegative integers k, the following equalities hold:*

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{A}_{\mathscr{k}}(x) = \frac{\mathcal{G}}{\mathcal{L} - x}\,\mathcal{A}_{\mathscr{k}}(x) + \frac{\mathrm{d}}{\mathrm{d}x}\mathcal{B}_{\mathscr{k}}(x), \tag{4.2}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{B}_{>\mathscr{k}}(x) = \mathcal{G}\biggl(\frac{\mathcal{L}}{\mathcal{L} - x} - \mathcal{B}_{\le \mathscr{k}-\mathcal{L}}(x)\biggr)\mathcal{B}_{>\mathscr{k}-\mathcal{L}}(x). \tag{4.3}$$

Here $\{B_{\le t}(x)\}$ is the sequence determined by the recurrence (3.3), $B_{\le -1}(x) := 0$, and $\mathcal{B}_{>-\mathcal{L}}(x) = \mathcal{B}_{\ge \mathscr{a}}(x)$ is the generating function of the expected numbers of leaves, i.e.

$$\mathcal{B}_{>-\mathcal{L}}(x) = \frac{x - \mathcal{L}}{\mathscr{G}} + \frac{\mathcal{L}}{\mathscr{G}(\mathcal{L} - x)^{\mathcal{G}}}.$$

Consequently,

$$f_k = 2\int_0^1 (1 - x)\mathcal{B}_{\mathscr{k}}(x)\,\mathrm{d}x. \tag{4.4}$$

Proof. Let 'root' denote the root of the random tree T_n on $[n]$. Let L_n denote the total number of leaves of T_n. For $n \ge 2$, $L_n = L' + L''$, where L' and L'', denote the total number of leaves in the left subtree T' and the right subtree T'', respectively. Let root$'$ (respectively, root$''$) denote the root of T' (respectively, T'') if this subtree is nonempty. If both subtrees are nonempty then

$$\mathbf{1}_{\{R(\text{root})>k)\}} = \mathbf{1}_{\{R(\text{root}')>k-1\}}\,\mathbf{1}_{\{R(\text{root}'')>k-1\}}, \qquad k \ge 0.$$

Let $0 < j < n - 1$. Now, conditioned on the event 'the vertex set of T' is a given set J of j elements from $[n] \setminus$ root', the subtrees T' and T'' are independent, and marginally distributed as T_j and T_{n-1-j}, respectively. So

$$\begin{aligned}
\mathbb{E}[\mathbf{1}_{\{R(\text{root})>k\}}\,L_n \mid J] &= \mathbb{E}[\mathbf{1}_{\{R(\text{root}')>k-1\}}\,\mathbf{1}_{\{R(\text{root}'')>k-1\}}(L' + L'') \mid J] \\
&= \mathbb{E}[\mathbf{1}_{\{R(\text{root}')>k-1\}}\,\mathbf{1}_{\{R(\text{root}'')>k-1\}}(L_j + L_{n-1-j})] \\
&= \mathbb{E}[\mathbf{1}_{\{R(\text{root}')>k-1\}}\,L_j]\,\mathbb{P}(R(\text{root}'') > k - 1) \\
&\quad + \mathbb{E}[\mathbf{1}_{\{R(\text{root}'')>k-1\}}\,L_{n-1-j}]\,\mathbb{P}(R(\text{root}') > k - 1) \\
&= f_{j,>k-1}\,p_{n-1-j,>k-1} + f_{n-1-j,>k-1}\,p_{j,>k-1},
\end{aligned}$$

where $p_{\nu,>k-1} := \mathbb{P}(R(\text{root of } T_\nu) > k - 1)$. Setting $f_{0,>k-1} = 0$, $p_{0,>k-1} = 1$, we see that the last equality also holds for $j = 0,\ n - 1$. Since $|J|$ is uniform on $\{0, \dots, n - 1\}$, we obtain

$$f_{n,>k} = \mathbb{E}[\mathbf{1}_{\{R(\text{root})>k\}}\,L_n] = \frac{2}{n}\sum_{j=0}^{n-1} f_{j,>k-1}\,p_{n-1-j,>k-1}.$$

It follows immediately that

$$\frac{\mathrm{d}}{\mathrm{d}x}\sum_{n\ge 1} f_{n,>k}x^n = 2\biggl(\sum_{n\ge 0} p_{n,>k-1}x^n\biggr)\biggl(\sum_{n\ge 1} f_{n,>k-1}x^n\biggr),$$

which is equivalent to (4.3), since

$$\sum_{n\ge 0} p_{n,>k-1}x^n = 1 + \sum_{n\ge 1}(1 - p_{n,\le k-1})x^n = 1 + \frac{x}{1 - x} - B_{\le k-1}(x) = \frac{1}{1 - x} - B_{\le k-1}(x).$$

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

Equation (4.2) is implied by a simple recurrence

$$F_{n,k} = f_{n,k} + \frac{2}{n}\sum_{j=0}^{n-1} F_{j,k}, \qquad n \geq 2,\ k \geq 0.$$

Finally, from (4.2), we obtain

$$f_k = \lim_{x\uparrow 1}(1-x)^2 \mathscr{A}_{\mathscr{k}}(\mathscr{x}) = \int_{\mathscr{a}}^{\mathscr{E}} (\mathscr{E} - x)^{\mathscr{G}} \frac{\mathrm{d}}{\mathrm{d}\mathscr{x}} \mathscr{B}_{\mathscr{k}}(\mathscr{x})\,\mathrm{d}\mathscr{x} = \mathscr{G}\int_{\mathscr{a}}^{\mathscr{E}} (\mathscr{E} - x)\mathscr{B}_{\mathscr{k}}(\mathscr{x})\,\mathrm{d}\mathscr{x}.$$

The proof of Lemma 4.1 is complete. □

Next, introduce

$$\widehat{A}_k(x) = \sum_{n\geq 1} x^n G_{n,k} \quad \text{and} \quad \widehat{B}_k(x) = \sum_{n\geq 1} x^n g_{n,k},$$

where $g_{n,k} := \mathbb{E}[\mathbf{1}_{\{R(\text{root})=k\}}\,\mathcal{L}_n]$ and $\mathcal{L}_n$ is the number of leaves closest to the root of the tree.

Lemma 4.2. *The following equalities hold:*

$$\frac{\mathrm{d}}{\mathrm{d}x}\widehat{A}_k(x) = \frac{2}{1-x}\widehat{A}_k(x) + \frac{\mathrm{d}}{\mathrm{d}x}\widehat{B}_k(x), \qquad k \geq 0, \tag{4.5}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\widehat{B}_k(x) = 2[1 + B_{\geq k-1}(x)]\widehat{B}_{k-1}(x), \qquad k > 0, \tag{4.6}$$

with $\widehat{B}_0(x) = x$. *Consequently,*

$$g_k = 2\int_0^1 (1-x)\widehat{B}_k(x)\,\mathrm{d}x. \tag{4.7}$$

Proof. Let us prove (4.6). Recall that $\mathcal{L}_n$ denotes the total number of leaves *closest* to the root of T_n. For $n \geq 2$, let $\mathcal{L}'$ and $\mathcal{L}''$ denote the total number of leaves in the left subtree T' and the right subtree T'' closest to the respective root. Let $k > 0$. If both subtrees are nonempty, i.e. $0 < j < n-1$ then

$$\begin{aligned}\mathbf{1}_{\{R(\text{root}=k)\}}\,\mathcal{L}_n &= \mathbf{1}_{\{R(\text{root}')=k-1\}}\,\mathbf{1}_{\{R(\text{root}'')=k-1\}}\,\mathcal{L}_n \\ &\quad + \mathbf{1}_{\{R(\text{root}')=k-1\}}\,\mathbf{1}_{\{R(\text{root}'')>k-1\}}\,\mathcal{L}_n + \mathbf{1}_{\{\mathscr{R}(\text{root}')>\mathscr{k}-\mathscr{E}\}}\,\mathbf{1}_{\{\mathscr{R}(\text{root}'')=\mathscr{k}-\mathscr{E}\}}\,\mathcal{L}_n \\ &= \mathbf{1}_{\{R(\text{root}')=k-1\}}\,\mathbf{1}_{\{R(\text{root}'')=k-1\}}(\mathcal{L}' + \mathcal{L}'') \\ &\quad + \mathbf{1}_{\{R(\text{root}')=k-1\}}\,\mathbf{1}_{\{R(\text{root}'')>k-1\}}\,\mathcal{L}' + \mathbf{1}_{\{\mathscr{R}(\text{root}')>\mathscr{k}-\mathscr{E}\}}\,\mathbf{1}_{\{\mathscr{R}(\text{root}'')=\mathscr{k}-\mathscr{E}\}}\,\mathcal{L}''.\end{aligned}$$

The contribution of the first product on the last right-hand side to $\mathbb{E}[\mathbf{1}_{\{R(\text{root})=k\}}\,\mathcal{L}_n \mid \mathcal{J}]$ is

$$g_{j,k-1}\,p_{n-1-j,k-1} + g_{n-1-j,k-1}\,p_{j,k-1}.$$

The total contribution of the second product and the third product is

$$g_{j,k-1}\,p_{n-1-j,>k-1} + g_{n-1-j,k-1}\,p_{j,>k-1},$$

so that

$$\mathbb{E}[\mathbf{1}_{\{R(\text{root})=k\}}\,\mathcal{L}_n \mid \mathcal{J}] = \mathscr{g}_{\mathscr{j},\mathscr{k}-\mathscr{E}}\,\mathscr{p}_{\mathscr{n}-\mathscr{E}-\mathscr{j},\geq\mathscr{k}-\mathscr{E}} + \mathscr{g}_{\mathscr{n}-\mathscr{E}-\mathscr{j},\mathscr{k}-\mathscr{E}}\,\mathscr{p}_{\mathscr{j},\geq\mathscr{k}-\mathscr{E}}.$$

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

The last equation continues to hold for $j = 0$ and $j = n - 1$, if we set $p_{0,\geq \ell} = 1$ for all $\ell \geq 0$. Consequently,

$$g_{n,k} = \mathbb{E}[\mathbf{1}_{\{R(\text{root})=k\}}\,\mathcal{L}_n] = \frac{\mathcal{G}}{n}\sum_{j=\mathcal{A}}^{n-\mathscr{E}} g_{j,k-\mathscr{E}}\,p_{n-\mathscr{E}-j,\geq k-\mathscr{E}},$$

and (4.6) follows immediately. And, as before, (4.5) is the direct consequence of

$$G_{n,k} = g_{n,k} + \frac{2}{n}\sum_{j=0}^{n-1} G_{j,k}, \qquad n \geq 2,\ k \geq 0.$$

Equation (4.7) is proved in the same way as (4.4). □

Introduce $\mathcal{L}_{n,k}$ and $\widehat{L}_{n,k}$, the total number of descendant leaves of rank-k vertices and the total number of descendant leaves *closest* to rank-k vertices. Recalling the notation $V_{n,k}$ for the total number of rank-k vertices, we see that $\mathcal{L}_{n,k}/\mathcal{V}_{n,k}$ and $\widehat{L}_{n,k}/V_{n,k}$ are the average numbers of descendant leaves and the closest descendant leaves per vertex of rank k.

Theorem 4.1. *For all nonnegative integers k, the following equalities hold:*

$$\lim_{n\to\infty}\mathbb{E}\biggl[\frac{\mathcal{L}_{n,k}}{V_{n,k}}\biggr] = \frac{f_k}{c_k}, \qquad \lim_{n\to\infty}\mathbb{E}\biggl[\frac{\widehat{L}_{n,k}}{V_{n,k}}\biggr] = \frac{g_k}{c_k}.$$

Proof. Consider $\mathcal{L}_{n,k}/\mathcal{V}_{n,k}$, for instance. Observe first that $\mathcal{L}_{n,k} \leq n$. Now, for $a = 1/k!$ and $\varepsilon > 0$, write

$$\begin{aligned}
\mathbb{E}\biggl[\frac{\mathcal{L}_{n,k}}{V_{n,k}}\biggr] &= \mathbb{E}\biggl[\frac{\mathcal{L}_{n,k}}{V_{n,k}}\,\mathbf{1}_{\{V_{n,k}<0.03an\}}\biggr] + \mathbb{E}\biggl[\frac{\mathcal{L}_{n,k}}{V_{n,k}}\,\mathbf{1}_{\{V_{n,k}\geq 0.03an\}}\,\mathbf{1}_{\{|V_{n,k}/n-c_k|>\varepsilon\}}\biggr] \\
&\quad + \mathbb{E}\biggl[\frac{\mathcal{L}_{n,k}}{V_{n,k}}\,\mathbf{1}_{\{V_{n,k}\geq 0.03an\}}\,\mathbf{1}_{\{|V_{n,k}/n-c_k|\leq\varepsilon\}}\biggr] \\
&= \mathbb{E}_1 + \mathbb{E}_2 + \mathbb{E}_3.
\end{aligned}$$

Here, by Lemma 2.4 and Corollary 2.2, respectively,

$$\mathbb{E}_1 \leq n\mathrm{e}^{-0.01an^{1-\delta}} \to 0, \qquad \mathbb{E}_2 = O\biggl(\mathbb{P}\biggl(\biggl|\frac{V_{n,k}}{n} - c_k\biggr| > \varepsilon\biggr)\biggr) \to 0 \quad \text{as } n \to \infty,$$

and

$$\begin{aligned}
\mathbb{E}_3 &= \frac{1}{n(c_k + O(\varepsilon))}\mathbb{E}[\mathcal{L}_{n,k}\,\mathbf{1}_{\{\mathcal{V}_{n,k}\geq\mathcal{A}.\mathcal{A}\mathcal{G}an\}}\,\mathbf{1}_{\{|\mathcal{V}_{n,k}/n-c_k|\leq\varepsilon\}}] \\
&= \frac{\mathbb{E}[\mathcal{L}_{n,k}/n]}{c_k + O(\varepsilon)}\biggl[1 + O\biggl(\mathbb{P}(V_{n,k} < 0.03an) + \mathbb{P}\biggl(\biggl|\frac{V_{n,k}}{n} - c_k\biggr| > \varepsilon\biggr)\biggr)\biggr].
\end{aligned}$$

Therefore,

$$\lim_{\varepsilon\downarrow 0}\limsup_{n\to\infty}\mathbb{E}_3 = \lim_{\varepsilon\downarrow 0}\liminf_{n\to\infty}\mathbb{E}_3 = \frac{f_k}{c_k}.$$

So $\lim_{n\to\infty}\mathbb{E}[\mathcal{L}_{n,k}/\mathcal{V}_{n,k}] = f_k/c_k$. The proof for $\mathbb{E}[\widehat{L}_{n,k}/V_{n,k}]$ is similar. □

Note that a slight modification of the proof of Corollary 2.2 shows that, in probability, $\mathcal{L}_{n,k}/n \to f_k$ and $\widehat{L}_{n,k}/n \to g_k$. Therefore, $\mathcal{L}_{n,k}/\mathcal{V}_{n,k} \to f_k/c_k$, and $\widehat{L}_{n,k}/V_{n,k} \to g_k/c_k$ in probability as well.

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

Using MAPLE® to integrate the differential equations (4.3) and (4.6), we compute $\{f_j\}_{j\le 2}$ and $\{g_j\}_{j\le 2}$ via (4.4) and (4.7), respectively, as

$$f_0 = \tfrac{1}{3}, \qquad f_1 = \tfrac{17}{30}, \quad f_2 = \tfrac{152\,389}{170\,100}, \qquad g_0 = \tfrac{1}{3}, \qquad g_1 = \tfrac{1}{3}, \qquad g_2 = \tfrac{49}{180}.$$

Therefore,

$$\frac{f_0}{c_0} = 1, \qquad \frac{f_1}{c_1} = \frac{17}{9}, \qquad \frac{f_2}{c_2} = \frac{152\,389}{361\,41}, \qquad \frac{g_0}{c_0} = 1, \qquad \frac{g_1}{c_1} = \frac{10}{9}, \qquad \frac{g_2}{c_2} = \frac{2205}{1721}.$$

Remark 4.1. (i) The fact that g_0, g_1 are both $\frac{1}{3}$ follows from the observation that, for $n \ge 2$, the number of pairs (v, u), where v is a rank-k vertex and u is its closest descendant leaf, is the same as the number of all leaves when $k = 0$ or $k = 1$.

(ii) The data suggest that both f_k/c_k and g_k/c_k increase with k, albeit at a slower rate for g_k/c_k.

5. Numerics and gap-free factorization conjecture

In conclusion we present some intriguing experimental data on number-theoretic properties of $\{c_k\}$. Recall that, by (3.4),

$$\sum_{j=0}^{k} c_j = 2\int_0^1 (1-y)B_{\le k}(y)\,\mathrm{d}y. \tag{5.1}$$

Then, using (3.3),

$$\begin{aligned}\int_0^1 (1-y)B_{\le k}(y)\,\mathrm{d}y &= \frac{1}{2}\int_0^1 (1-y)^2 B_{\le k}(y)'\,\mathrm{d}y\\ &= \frac{1}{2}\int_0^1 [2-2y+y^2-(1-(1-y)B_{\le k-1}(y))^2]\,\mathrm{d}y,\end{aligned}$$

so

$$\sum_{j=0}^{k} c_j = \int_0^1 [2-2y+y^2-[1-(1-y)B_{\le k-1}(y)]^2]\,\mathrm{d}y. \tag{5.2}$$

Equation (5.2) enables us to compute c_k directly through $B_{\le k-1}(x)$, without knowing $B_k(x)$.

Using this simplification, we have obtained the exact values of c_4 and c_5. That is, we have computed that c_4 is equal to

$$\frac{122\,058\,464\,141\,653\,662\,196\,290\,113\,232\,646\,304\,412\,999\,902\,283\,512\,425\,580\,156\,787\,323}{3\,353\,377\,025\,022\,449\,199\,852\,900\,725\,670\,960\,067\,418\,280\,803\,797\,231\,788\,288\,000\,000\,000},$$

a fraction whose denominator has 67 digits, and whose approximate value is 0.0364. Combining this with the values of c_i for $i \le 4$, we see that about 99.14 percent of all vertices in all trees of size n are of rank 4 or less, and the same holds with high probability for the random tree as well.

The prime factorization of the denominator $\operatorname{denom}(c_4)$, when c_4 is written in simplest terms, obtained by MAPLE, is even more interesting, since it is

$$\operatorname{denom}(c_4) = 2^{17} \times 3^{18} \times 5^9 \times 7^8 \times 11^8 \times 13^7 \times 17^6 \times 19^5 \times 23^4 \times 29^2 \times 31.$$

So the largest prime divisor of $\operatorname{denom}(c_4)$ is 31, which is a tiny number compared to $\operatorname{denom}(c_4)$. Even more striking is the fact that $\operatorname{denom}(c_4)$ is divisible by *every prime* up to 31. In stark

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

contrast, the numerator of c_4, while comparable in size to the denominator, is the product of just *two* primes, the smaller of which is 232 196 467.

This surprising fact warrants a second look at the numbers c_k for $k \leq 3$, already computed in [2]. The factorized representation of the denominators, including $\operatorname{denom}(c_4)$ are as follows:

- $\operatorname{denom}(c_0) = 3$,
- $\operatorname{denom}(c_1) = 2 \times 5$,
- $\operatorname{denom}(c_2) = 2^2 \times 3^4 \times 5^2$,
- $\operatorname{denom}(c_3) = 2^8 \times 3^7 \times 5^5 \times 7^3 \times 11^3 \times 13^2 \times 17$, and
- $\operatorname{denom}(c_4) = 2^{17} \times 3^{18} \times 5^9 \times 7^8 \times 11^8 \times 13^7 \times 17^6 \times 19^5 \times 23^4 \times 29^2 \times 31$.

So all $\operatorname{denom}(c_k)$ for $k \leq 4$ have very small prime divisors. With the exception of $k = 0$ and $k = 1$, it seems that the prime divisors of $\operatorname{denom}(c_k)$ are *precisely* the first t prime numbers for some t. Those two exceptions may be a reflection of how relatively simple the counting of leaves and their fathers is.

Even though the computation of c_4 was already exceptionally time consuming, we decided to compute the next value c_5. This task turned out to be so problematic that time and again we were tempted to give up. Mobilizing all the insight into the algebraic form of the functions $B_j(x)$, we eventually obtained the answer. The approximate value of c_5 is 0.0074. So, with high probability, about 99.875 percent of all vertices are of rank 5 or less. The number $\operatorname{denom}(c_5)$ has 274 digits, and its prime factorization is

$$2^{48} \times 3^{42} \times 5^{28} \times 7^{18} \times 11^{16} \times 13^{16} \times 17^{17} \times 19^{16} \times 23^{15} \times 29^{12} \times 31^{12} \times 37^{10} \times 41^9 \\ \times 43^8 \times 47^7 \times 53^5 \times 59^3 \times 61^2.$$

If not for this strikingly simple factorization, we would not dare to type in the 274-digit long monster. So yet again, $\operatorname{denom}(c_k)$ has only very small prime factors, and it is divisible by every prime up to its largest prime factor, 61. (As for the numerator, its smallest prime divisor must be extremely large as MAPLE's factorization algorithm failed the task.)

Based on these data points, we guessed at and proved the following theorem.

Theorem 5.1. *Let* $\operatorname{denom}(c_k)$ *be the denominator of* c_k *when* c_k *is written in its smallest terms. Then the largest prime divisor of the denominator is at most* $2^{k+1} + 1$*; this bound is attained for* $k = 3, 4, 5$.

Conjecture 5.1. *Let* $k \geq 2$*, and let* p_k *be largest prime divisor of* $\operatorname{denom}(c_k)$. *Then* $\operatorname{denom}(c_k)$ *is divisible by every prime less than* p_k.

Perhaps it is also true that the smallest prime divisor of the numerator of c_k grows super-exponentially with k, but we hesitate to make any specific guess. The reason the second conjecture is out of reach for now is simple: the numerator of c_k is a sum of a very large set of summands, and we are unable to prove that the sum will not be divisible by at least as high a power of a given prime p as the denominator of c_k.

Proof of Theorem 5.1. Before embarking on the proof, we will need a few simple technical lemmas.

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

Recall that $B_k(x)$ denotes the exponential generating function for the numbers of trees on vertex set $[n]$ whose root is of rank k. The first two examples are $B_0(x) = x$, and $B_1(x) = 2\log(1/(1-x)) - 2x - x^3/3$.

Lemma 5.1. *For all natural numbers k, we have $B_k(x) \in$ **PL**, meaning that $B_k(x)$ is a bivariate polynomial $\mathbb{P}_k(u, v)$ at $u = (1-x)$, $v = \log 1/(1-x)$.*

Proof. See Lemma 4.1 of [2]. □

It was also proved in [2] that the class **PL** is closed under integration. In fact, the following, stronger statement is true.

Lemma 5.2. *Let b and c be nonnegative integers, and let us write*

$$\int (1-x)^b \log\left(\frac{1}{1-x}\right)^c \mathrm{d}x = \sum_{i=1}^{m} a_i (1-x)^{b_i} \log\left(\frac{1}{1-x}\right)^{c_i}$$

with the rational numbers a_i written in their simplest form. Then, for all i, the denominator of a_i has no prime divisor larger than $b+1$.

Proof. This follows by induction on c, the initial case of $c = 0$ being obvious. Indeed, integration by parts yields

$$\begin{aligned}\int (1-x)^b \log\left(\frac{1}{1-x}\right)^c \mathrm{d}x \\ = -\log\left(\frac{1}{1-x}\right)^c \frac{(1-x)^{b+1}}{b+1} + \int \frac{(1-x)^b}{b+1} c \log\left(\frac{1}{1-x}\right)^{c-1} \mathrm{d}x,\end{aligned} \tag{5.3}$$

and the proof is complete. Note that this argument also shows that in the statement of Lemma 5.2, the inequalities $b_i \leq b+1$ and $c_i \leq c$ hold. □

Note that equation (5.3) implies that

$$I_{b,c} := \int_0^1 (1-x)^b \log\left(\frac{1}{1-x}\right)^c \mathrm{d}x = \frac{\mathbf{1}_{\{c=0\}}}{b+1} + \frac{c}{b+1} I_{b,c-1},$$

so iterating the same operation, we obtain

$$I_{b,c} = \frac{c!}{(b+1)^{c+1}}. \tag{5.4}$$

Lemma 5.3. *When written in simplest form, no term of $B_k(x)$ has a denominator with a prime divisor larger than $2^{k+1} - 1$. Furthermore, both the exponent b_i of $(1-x)$ and the exponent c_i of $\log(1/(1-x))$ in the* **PL** *form of $B_k(x)$ are at most as large as $2^{k+1} - 1$.*

Proof. We prove the lemma by strong induction on k. It is straightforward to check that $B_0(x)$ and $B_1(x)$ satisfy both requirements. Now let us assume that the claims of the lemma are true for all $B_j(x)$ with $j < k$, and prove them for B_k. Equation (3.1) shows that $B'_k(x)$ is a quadratic form of $B_i(x)$ with $i < k$ and $(1-x)^{-1}$. Consequently, $B'_k(x)$ is of the form $\sum_{i=1}^m a_i (1-x)^{b_i} \log(1/(1-x))^{c_i}$, where $b_i \geq -1$ is an integer, while a_i is rational and c_i is a nonnegative integer. Moreover, it follows from (3.1) and the induction hypothesis that, in the

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

sum representing $B_k'(x)$, both the exponent b_i of $(1-x)$ and the exponent c_i of $\log(1/(1-x))$ are at most as large as $2(2^k-1)=2^{k+1}-2$.

Now the contribution of $\sum_{i:b_i=-1} a_i(1-x)^{b_i}\log(1/(1-x))^{c_i}$ to $B_k(x)$ itself is

$$\sum_{i:b_i=-1}\frac{a_i}{c_i+1}\log\biggl(\frac{1}{1-x}\biggr)^{c_i+1}$$

with $c_i+1\le 2^{k+1}-1$. As for the contribution to $B_k(x)$ of the remaining summands with $b_i\ge 0$, using Lemma 5.2 and by (5.3), we see that in all the summands neither the exponent of $(1-x)$ nor the exponent of $\log(1/(1-x))$ can exceed $2^{k+1}-1$, since integration of the terms with $b_i\ge 0$ and $c_i\ge 0$ will increase these exponents by at most 1. As addition and multiplication of terms will not result in the appearance of a larger prime divisor, the claim for $B_k(x)$ is proved. □

Armed with these lemmas, we complete the proof of Theorem 5.1 as follows. By (5.1),

$$c_k=\lim_{x\uparrow 1}(1-x)^2A_k(x)=2\int_0^1(1-x)B_k(x)\,\mathrm{d}x.$$

Here

$$B_k(x)=\sum_i a_i(1-x)^{b_i}\biggl(\log\frac{1}{1-x}\biggr)^{c_i},\qquad 0\le b_i,\ c_i\le 2^{k+1}-1,$$

and no a_i has a denominator with a prime divisor larger than $2^{k+1}-1$. From (5.4), it follows that c_k is the sum of rational numbers, whose denominators do not have prime divisors exceeding $2^{k+1}+1$, which is a common upper bound for the largest denominator of a_i and for the largest b_i+2. This completes the proof of the theorem. □

Acknowledgements

The authors are thankful to Frank Garvan who on numerous occasions advised them on how to convince Maple to carry out a difficult task. We owe a debt of gratitude to Svante Janson for encouraging and helpful feedback on the original manuscript. An anonymous referee read the paper with meticulous attention to technical details, pinpointing, in particular, two omissions that had to be fixed. Thank you.

References

[1] Aldous, D. (1991). Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Prob.* **1,** 228–266.

[2] Bóna, M. (2014). k-protected vertices in binary search trees. *Adv. Appl. Math.* **53,** 1–11.

[3] Devroye, L. (1986). A note on the height of binary search trees. *J. Assoc. Comput. Mach.* **33,** 489–498.

[4] Devroye, L. (1991). Limit laws for local counters in random binary search trees. *Random Structures Algorithms* **2,** 303–315.

[5] Devroye, L. and Janson, S. (2014). Protected nodes and fringe subtrees in some random trees. *Electron. Commun. Prob.* **19,** 6.

[6] Drmota, M. (2009). *Random Trees: An Interplay Between Combinatorics and Probability*. Springer, Vienna.

[7] Du, R. R. X. and Prodinger, H. (2012). Notes on protected nodes in digital search trees. *Appl. Math. Lett.* **25,** 1025–1028.

[8] Flajolet P. and Sedgewick, R. (2009). *Analytic Combinatorics*. Cambridge University Press.

[9] Fuchs, M., Lee, C.-K. and Yu, G.-R. (2016). On 2-protected nodes in random digital trees. *Theoret. Comput. Sci.* **622,** 111–122.

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press

[10] Holmgren, C. and Janson, S. (2015). Asymptotic distribution of two-protected nodes in ternary search trees. *Electron J. Prob.* **20,** 9.

[11] Kesten, H. and Pittel, B. (1996). A local limit theorem for the number of nodes, the height, and the number of final leaves in a critical branching process tree. *Random Structures Algorithms* **8,** 243–299.

[12] Kolchin, V. F. (1978). Moment of degeneration of a branching process and height of a random tree. *Math. Notes Acad. Sci. USSR* **24,** 954–961.

[13] Mahmoud, H. and Pittel, B. (1984). On the most probable shape of a search tree grown from a random permutation. *SIAM J. Algebraic Discrete Meth.* **5,** 69–81.

[14] Mahmoud, H. M. and Ward, M. D. (2012). Asymptotic distribution of two-protected nodes in random binary search trees. *Appl. Math. Lett.* **25,** 2218–2222.

[15] Mahmoud, H. M. and Ward, M. D. (2015). Asymptotic properties of protected notes in random recursive trees. *J. Appl. Prob.* **52,** 290–297.

[16] Pitman, J. (2006). *Combinatorial Stochastic Processes* (Lecture Notes Math. **1875**). Springer, Berlin.

[17] Pittel, B. (1984). On growing random binary trees. *J. Math. Anal. Appl.* **103,** 461–480.

[18] Pittel, B. (1994). Note on the heights of random recursive trees and random m-ary search trees. *Random Structures Algorithms* **5,** 337–347.

https://doi.org/10.1017/apr.2017.24 Published online by Cambridge University Press