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

MIKLÓS BÓNA AND BORIS PITTEL

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 $k$. Notoriously hard to compute, the exact fractions $c_k$ had been determined for $k \leq 3$ only. We computed $c_4$ and $c_5$ as well; both are ratios of enormous integers, denominator of $c_5$ being 274 digits long. Prompted by the data, we proved that, in sharp contrast, the largest prime divisor of $c_k$'s denominator is $2^{k+1}+1$ at most. 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 $2^{k+1}+1$.

## 1. Introduction

1.1. **Background and Definitions.** Various parameters of many models of random rooted trees are fairly well understood *if they relate to a near-root part of the tree or to 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 Flajolet and Sedgewick [6] for a comprehensive treatment of these results. The tree height and width are parameters of global nature, see Kolchin [8], Devroye [3], Mahmoud and Pittel [9], Pittel [12], Kesten and Pittel [7], Pittel [13], for instance. 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 Aldous [1], Mahmoud and Ward [10], [11], Bóna [2], and Devroye and Janson [4]. Diversity of models and techniques notwithstanding, a salient feature of these studies is usage of inherently recursive nature of the random trees in question. Deletion of the tree root produces a forest of rooted subtrees that are conditionally independent, each being distributed as the random tree for the properly chosen tree size.

*Date*: December 14, 2014.

2010 *Mathematics Subject Classification.* 05A05, 05A15, 05A16, 05C05, 06B05, 05C80, 05D40, 60C05.

*Key words and phrases.* search tree, root, leaves, ranks, enumeration, asymptotic, distribution, numerical data.

Not surprisingly, the technical details of fringe analysis become quite complex as soon soon 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, remained an open problem. In the present 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 $[n] = \{1, 2, \cdots, n\}$ 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 an 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 labeled $n$, 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 labeled $a$, and the right one labeled $b$. If $n$ is the first (resp. last) entry of $p$, then the root will have only one child, and that is a left (resp. right) child, and it will necessarily be labeled $n-1$ 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 [5], [10] concerning the neighbors of leaves, Miklós 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 infinity. He also computed that $c_0 = 1/3$, $c_1 = 3/10$, $c_2 = 1721/8100$, and $c_3 \approx 0.105$. It is worth mentioning that a few months later, Svante Janson and Luc Devroye computed the same four values of $c_k$ with a completely different method. (The numbers $c_k$ are completely determined *theoretically*, but progressively more difficult to compute as $k$ increases.) These data show that roughly 95.5 percent of all vertices are of rank at most three, and raises the very intriguing questions whether $\{c_k\}$ is a probability distribution, and if yes, whether it is the limiting distribution of the rank of the uniformly random vertex of the tree. We were also keen to find a way for precise evaluation of the next constants, $c_4$ and $c_5$ at least.

1.3. **Main results.** In this paper, we are able to answer these questions. Here are our main results.

**Theorem 1.1. (i)** *The equality* $\sum_{k\ge 0} c_k = 1$ *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* $0 < \rho < 3/2$, $\lim_{n\to\infty} E\big[\rho^{R_n}\big] = E\big[\rho^{R}\big] < \infty$. *Consequently* $R_n \to R$ *in distribution, and with all its moments, and* $c_k = O(q^k)$ *for every* $0 < q < 2/3$. **(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* $R$.

The 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.

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

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

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

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

We also found a way to simplify computation of the numbers $c_k$ which enabled us to obtain the precise values of $c_4$ and $c_5$, thus going beyond $c_0, \dots, c_3$ determined in [2] and [4]. Our numerical results show that, with high probability, about 99.875 percent of all vertices are of rank five or less. When written in simplest form, the numerators and denominators of the rational numbers $c_k$ grow very fast. For instance, the denominator of $c_5$ ($denom(c_5)$) has 274 digits. Despite its enormity, the largest prime divisor of $denom(c_5)$ is 61. We conjectured and proved that this remarkable pattern holds for all $k$: the largest prime divisor of $denom(c_k)$ 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 $k \ge 2$, the set of prime divisors of $denom(c_k)$ 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 $k$.

## 2. Convergence of the random rank $R_n$

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

$$c_k = \lim_{n\to\infty} \frac{E_{n,k}}{n}, \quad k \ge 0.$$

Equivalently, $c_k$ is the limiting probability that $R_n$, the rank of the 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 $\mathrm{P}(R = k) = c_k$ and $R_n \Rightarrow R$ in distribution. Our first theorem confirms this conjecture and shows that the moment generating function of $R_n$ converges to that of $R$ for any argument below $3/2$.

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

*Proof.* Let $p_{n,k}$ be the probability that the root is of rank $k$. Then, for $n > 1$,

$$E_{n,k} = p_{n,k} + \frac{1}{n}\sum_{j=0}^{n-1}\left(E_{j,k} + E_{n-1-j,k}\right), \tag{1}$$

Indeed, the above formula just adds the expected value of indicator of the event "root is of rank $k$" to the expected total count of the non-root 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 $c_k := \lim E_{n,k}/n$, rational or not, will follow immediately from the next lemma.

**Lemma 2.2.** *Let $\{x_n\}$, $y_n$ be such that $y_n = O(n^{1-\varepsilon})$, $(\varepsilon > 0)$, and*

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

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

*Proof.* First of all, (1) is equivalent to

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

Standard manipulation shows then that

$$nx_n - (n+1)x_{n-1} = ny_n - (n-1)y_{n-1}, \quad n > 1, \tag{2}$$

or

$$\begin{aligned}\frac{x_n}{n+1} - \frac{x_{n-1}}{n} &= \frac{y_n}{n+1} - \frac{y_{n-1}}{n}\frac{n-1}{n+1}\\ &= \frac{y_n}{n+1} - \frac{y_{n-1}}{n} + O\big(n^{-1-\varepsilon}\big).\end{aligned}$$

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 $\lim_{n\to\infty} x_n/(n+1)$, and so does $\lim_{n\to\infty} x_n/n$. □

Since $p_{n,k} = O(1)$, the conditions of Lemma 2.2 obviously hold for $x_n = E_{n,k}$ and $y_n = p_{n,k}$ with $\varepsilon \in (0, 1]$. Consequently, for each $k \geq 0$, there exists a finite limit $c_k := \lim E_{n,k}/n$. Further,

$$\sum_k \frac{E_{n,k}}{n} = 1 \Longrightarrow \sum_k c_k \leq 1. \tag{3}$$

Next, given $\rho > 1$, introduce

$$\mathcal{H}_n(\rho) = \sum_{k \leq n-1} \rho^k E_{n,k},$$

the expected value of $\sum_{v \in [n]} \rho^{R(v)}$, $R(v)$ denoting the rank of a generic vertex $v$. Then, analogously to (1),

$$\mathcal{H}_n(\rho) = h_n(\rho) + \frac{1}{n} \sum_{j=0}^{n-1} \left( \mathcal{H}_j(\rho) + \mathcal{H}_{n-1-j}(\rho) \right), \quad n > 1, \tag{4}$$

where $h_n(\rho) = \mathrm{E}\left[\rho^{R(root)}\right]$. 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

$$\rho^{R(root)} \leq \frac{\sum_j \rho^j X_{n,j}}{L_n} \Longrightarrow h_n(\rho) \leq \mathrm{E}\left[ \frac{\sum_j \rho^j X_{n,j}}{L_n} \right]. \tag{5}$$

We will show that $L_n$ is of order $n$ so it is likely that $h_n(\rho)$ is at most of order $n^{-1} \sum_j \rho^j \mathrm{E}[X_{n,j}]$. So let us bound $\sum_j \rho^j \mathrm{E}[X_{n,j}]$. To do so, attach to the random tree "external" vertices so that every vertex of the tree itself has exactly two descendants; thus every leaf $\ell$ gets two external descendants, and every non-leaf 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 [9] that

$$\mathcal{L}_j(x) := \sum_{n \geq 1} \mathrm{E}[\mathcal{X}_{n,j}] x^n = \frac{2^j}{j!} \left( \log \frac{1}{1-x} \right)^j, \quad j > 0. \tag{6}$$

Introduce $L_j(x) = \sum_{n \geq 0} x^n \mathrm{E}[X_{n,j}]$; so $L_0(x) = x$. Arguing as in [9], it can be shown that, for $j \geq 2$,

$$\frac{dL_j(x)}{dx} = \frac{2}{1-x} L_{j-1}(x).$$

Notice that $[x^n] L_0(x) \leq [x^n] \log \frac{1}{1-x}$ for every $n \geq 0$. By induction on $j$, it follows that, for $j > 0$,

$$\mathrm{E}[X_{n,j}] = [x^n] L_j(x) \leq \frac{2^j}{j!} [x^n] \left( \log \frac{1}{1-x} \right)^j = \mathrm{E}[\mathcal{X}_{n,j}]. \tag{7}$$

Therefore, for every $r > 0$,

$$\begin{aligned}\sum_j r^j \mathrm{E}[X_{n,j}] &= [x^n] \sum_{j\ge 0} r^j L_j(x) \le [x^n] \sum_{j\ge 0} r^j \mathcal{L}_j(x)\\ &= [x^n] \sum_{j\ge 0} \frac{(2r)^j}{j!} \left(\log \frac{1}{1-x}\right)^j = [x^n] \exp\left[2r \log \frac{1}{1-x}\right]\\ &= [x^n](1-x)^{-2r} = \binom{n+2r-1}{n} = \frac{\Gamma(n+2r)}{\Gamma(n)\Gamma(2r)}\\ &= O\big(n^{2r-1}\big),\end{aligned}$$

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

$$\sum_j r^j \mathrm{E}[X_{n,j}] = O\big(n^{2r-1}\big). \tag{8}$$

Consequently for the numerator in the bound (5) of $h_n(\rho)$ we have

$$\mathrm{E}\left[\sum_j \rho^j X_{n,j}\right] = O\big(n^{2\rho-1}\big).$$

It remains to show that the denominator $L_n$ in (5) is quite likely to be of order $n$, so that $h_n(\rho) = O\big(n^{2\rho-1}/n\big) = O\big(n^{2\rho-2}\big)$. To be more specific, since $\mathrm{E}[L_n] = (n+1)/3$, [2], we should expect that $\mathrm{P}(L_n < an)$ is very small if $a < 1/3$.

**Lemma 2.3.** *If $x \in (0,1]$ and $y \in (0, y(x))$,*

$$y(x) := (2\sqrt{1-x})^{-1} \log \tfrac{1+\sqrt{1-x}}{1-\sqrt{1-x}}, \tag{9}$$

*then, setting $L_0 = 0$,*

$$\sum_{n\ge 0} y^n E\big[x^{L_n}\big] = \sqrt{1-x}\, \frac{1 + e^{2\sqrt{1-x}y} \frac{1-\sqrt{1-x}}{1+\sqrt{1-x}}}{1 - e^{2\sqrt{1-x}y} \frac{1-\sqrt{1-x}}{1+\sqrt{1-x}}}. \tag{10}$$

*Proof.* Since for $n > 1$

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

we obtain

$$\begin{aligned}\frac{\partial}{\partial y} \sum_{n\ge 0} y^n \mathrm{E}\big[x^{L_n}\big] &= x + \sum_{n\ge 2} y^{n-1} \sum_{k=0}^{n-1} \mathrm{E}\big[x^{L_k}\big] \mathrm{E}\big[x^{L_{n-1-k}}\big]\\ &= \left(\sum_{n\ge 0} y^n \mathrm{E}\big[x^{L_n}\big]\right)^2 - (1-x).\end{aligned} \tag{11}$$

Integrating and using $\sum_{n\ge 0} y^n \mathrm{E}\big[x^{L_n}\big]\Big|_{y=0} = 1$, we obtain (10), provided that the denominator in (10) is positive, a condition equivalent to $y < y(x)$. □

**Corollary 2.4.** *Let $a < 1/3$. For $\delta \in (0,1)$,*

$$P(L_n < an) \le \exp\big(-(1/3-a)n^{1-\delta}/2\big).$$

*Proof.* We start with a Chernoff-type bound

$$\mathrm{P}(L_n < an) \le x^{-an} y^{-n} \sum_{\nu\ge 0} y^\nu \mathrm{E}\big[x^{L_\nu}\big], \quad \forall\, x<1,\, y<y(x). \tag{12}$$

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

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

so we may choose $y = \exp\big(bn^{-\delta}\big)$, $b = (a+1/3)/2$. Using (12) and (10), it follows that

$$\mathrm{P}(L_n < an) = O\big[\exp(an^{1-\delta} - bn^{1-\delta})\big] = O\big[\exp(-(1/3-a)n^{1-\delta}/2\big].$$

□

Armed with the corollary, we return to (5). By Cauchy-Schwartz inequality,

$$\sum_j \rho^j X_{n,j} = \sum_\ell \rho^{|\mathcal{P}(\ell)|} \le X_n^{1/2} \left(\sum_\ell \rho^{2|\mathcal{P}(\ell)|}\right)^{1/2} \le n^{1/2} \left(\sum_j \rho^{2j} X_{n,j}\right)^{1/2}.$$

Therefore, applying Cauchy-Schwartz inequality again and using (8),

$$\begin{aligned}
\mathrm{E}\left[\mathbf{1}_{\{X_n\le an\}} \sum_j \rho^j X_{n,j}\right] &\le n^{1/2} \left(\mathrm{E}\big[\mathbf{1}_{X_n\le an}\big]\right)^{1/2} \left(\mathrm{E}\left[\sum_j \rho^{2j} X_{n,j}\right]\right)^{1/2} \\
&= n^{1/2}\mathrm{P}^{1/2}(X_n \le an) \left(\sum_j \rho^{2j}\mathrm{E}[X_{n,j}]\right)^{1/2} \\
&= O\big[n^{1/2} n^{(2\rho^2-1)/2}\mathrm{P}^{1/2}(X_n \le an)\big] \\
&= O\big[n^{\rho^2}\mathrm{P}^{1/2}(X_n \le an)\big]
\end{aligned}$$

Using the bound (8) with $\rho^2$ instead of $\rho$ and Corollary 2.4, we obtain then

$$\mathrm{E}\left[\mathbf{1}_{\{X_n\le an\}} \sum_j \rho^j X_{n,j}\right] = O\big(n^{\rho^2} \exp(-(1/3-a)n^{1-\delta}/4)\big) = o(1).$$

Therefore, by (5) and (8),

$$h_n(\rho) \le \mathrm{E}\left[\mathbf{1}_{\{X_n<an\}}\sum_j \rho^j X_{n,j}\right] + \frac{1}{an}\sum_j \rho^j \mathrm{E}[X_{n,j}] \\ = o(1) + O\big(n^{2\rho-2}\big). \tag{13}$$

**Lemma 2.5.** *For every fixed $\rho < 3/2$, there exists a finite $\lim_{n\to\infty} n^{-1}\mathcal{H}_n(\rho)$. 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 (8) and (13), $x_n := \mathcal{H}_n(\rho)$ and $y_n := h_n(\rho)$ satisfy the condition of Lemma 2.2 with $\varepsilon \in (0, 3-2\rho)$. Hence, there exists a finite

$$\lim_{n\to\infty} n^{-1}\mathcal{H}_n(\rho) = \lim_{n\to\infty} n^{-1}\mathrm{E}\left[\sum_{v\in[n]} \rho^{R(v)}\right] = \lim_{n\to\infty} n^{-1}\sum_{k\le n-1}\rho^k E_{n,k}.$$

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

$$\lim_{n\to\infty} n^{-1}\sum_{0<k\le n-1}\rho^k E_{n,k} = \sum_{k\ge 0}\rho^k c_k < \infty.$$

□

From Lemma 2.5 it follows that $R_n$, the rank $R(v)$ of the uniformly random vertex $v$, converges in distribution to $R$, ($\mathrm{P}(R=k) = c_k$, $k \ge 0$) fast enough for $\mathrm{E}[\rho^{R_n}]$ to converge to $\mathrm{E}[\rho^R]$ if $\rho < 3/2$. 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 2.6.** *Let $t > 1$ be fixed. For an ordered, fixed, $t$-tuple $\mathbf{k} = (k_1, \dots, k_t)$, let $p_n(\mathbf{k})$ denote the probability that the uniformly random $t$-tuple of vertices $(v_1, \dots, v_t)$ have ranks $R(v_1) = k_1, \dots, R(v_t) = k_t$. Then $\lim_{n\to\infty} p_n(\mathbf{k}) = \prod_{j=1}^t c_{k_j}$.*

*Proof.* For brevity, we consider $t = 2$ only. Let $E_{n,\mathbf{k}}$ denote the expected number of ordered pairs of vertices with ranks $k_1$ and $k_2$ respectively; so $E_{n,\mathbf{k}} = n(n-1)p_n(\mathbf{k})$. Then

$$E_{n,\mathbf{k}} = E'_{n,\mathbf{k}} + E''_{n,\mathbf{k}};$$

here $E'_{n,\mathbf{k}}$ is the contribution of the ordered pairs $(v_1, v_2)$ such that $v_1$ is not a descendant of $v_2$, and $v_2$ is not a descendant of $v_1$. $E''_{n,\mathbf{k}}$ comes from the remaining pairs $(v_1, v_2)$. Obviously $E''_{n,\mathbf{k}} \le 2\mathcal{E}_n$, $\mathcal{E}_n$ being the expected number of pairs $(v_1, v_2)$ such that $v_2$ isa descendant of $v_1$. Then, for $n > 1$,

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

Therefore $E''_{n,\mathbf{k}} = O(n \log n)$. Turn to $E'_{n,\mathbf{k}}$. This time

$$E'_{n,\mathbf{k}} = \frac{2}{n} \sum_{j=0}^{n-1} E_{j,k_1} E_{n-j-1,k_2} + \frac{2}{n} \sum_{j=0}^{n-1} E'_{j,\mathbf{k}};$$

the first sum accounts for pairs $(v_1, v_2)$ such that $v_1$ and $v_2$ do not belong to the same subtree, whence the product $E_{j,k_1} E_{n-j-1,k_2}$ of the expected (conditional) counts of vertices of rank $k_1$ and of rank $k_2$, in the left subtree and the right subtree respectively. We know that, for a fixed $k$, $E_{\nu,k} = \nu c_k + o(\nu)$ if $\nu \to \infty$. It follows then easily that

$$\frac{2}{n} \sum_{j=0}^{n-1} E_{j,k_1} E_{n-j-1,k_2} = c_{k_1} c_{k_2} \frac{n^2}{3} + o(n^2).$$

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

$$\frac{2}{n} \sum_{j=0}^{n-1} E_{j,k_1} E_{n-j-1,k_2} \le b_n^+ := \frac{n^2}{3} c_{k_1} c_{k_2} + \varepsilon n^2 + A. \tag{14}$$

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

$$\mathcal{E}^+_{n,\mathbf{k}} = b_n^+ + \frac{2}{n} \sum_{j=0}^{n-1} \mathcal{E}^+_{j,\mathbf{k}}, \quad \mathcal{E}^+_{j,\mathbf{k}} = 0, \ (j = 0, 1).$$

So, as usual,

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

here, using (14),

$$\frac{j b_j^+ - (j-1) b_{j-!}^+}{j(j+1)} = \frac{(c_{k_1} c_{k_2} + 3\varepsilon) j^2 + O(j)}{j^2 + O(j)} = c_{k_1} c_{k_2} + 3\varepsilon + O(j-1).$$

Consequently

$$\mathcal{E}^+_{n,\mathbf{k}} = \big[c_{k_1} c_{k_2} + 3\varepsilon\big] n^2 + O(n \log n).$$

This implies

$$\limsup \frac{E'_{n,\mathbf{k}}}{n(n-1)} \le \lim \frac{\mathcal{E}^+_{n,\mathbf{k}}}{n^2} = c_{k_1} c_{k_2} + 3\varepsilon.$$

Analogously,

$$\liminf \frac{E'_{n,\mathbf{k}}}{n(n-1)} \ge c_{k_1} c_{k_2} - 3\varepsilon.$$

Letting $\varepsilon \downarrow 0$, we obtain $\lim \frac{E'_{n,\mathbf{k}}}{n(n-1)} = c_{k_1} c_{k_2}$. Since $E''_{n,\mathbf{k}} = O(n \log n)$, we conclude that $\lim \frac{E_{n,\mathbf{k}}}{n(n-1)} = c_{k_1} c_{k_2}$. □

**Corollary 2.7.** *Introduce $V_{n,k}$, the total number of vertices of rank $k$; so $V_{n,0} = L_n$, the total number of leaves. Then $V_{n,k}/n \to c_k$ in probability, i. e. for every $\varepsilon > 0$, $P(|V_{n,k}/n - c_k| > \varepsilon) = o(1)$ as $n \to \infty$.*

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

That $o(1)$ in the Corollary would not be enough for us. Recall though that for $L_n := V_{n,0}$ we were able to show (Corollary 2.4) that $V_{n,0} < (c_0-\varepsilon)n$ with probability $\exp(-\varepsilon n^{1-\delta})$ at most, 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 2.8.** *For $\delta < 1$ and $n$ large enough,*

$$P(V_{n,k} < 0.03an) \le \exp\big(-0.01an^{1-\delta}\big), \quad a := 1/k!.$$

*Proof.* **(i)** Clearly $V_{n,k} \ge \mathcal{V}_{n,k}$, which is the total number of vertex-to-leaf paths of length $k$ such that every non-leaf vertex of the path has only one child. Introduce $F(x,y) = \sum_{n\ge 0} y^n \mathrm{E}\big[x^{\mathcal{V}_{n,k}}\big]$, ($\mathcal{V}_{0,k} := 0$). For $y < 1$, $F(1,y) = (1-y)^{-1}$; so for $x \le 1$, $y < 1$, we have $F(x,y) \le (1-y)^{-1} < \infty$.

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

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

It follows after simple algebra that

$$\frac{\partial}{\partial y}F(x,y) = F^2(x,y) - (1-x)y^k\frac{2^k}{k!}, \tag{15}$$

blending with (11) for $k = 0$. Consequently, for $y \ge 1/2$,

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

Introduce $G(x,y)$, $(y \ge 1/2)$, the solution of

$$\frac{\partial}{\partial y}G(x,y) = G^2(x,y) - a(1-x), \quad G(x,1/2) = F(x,1/2).$$

Integrating the last equation and using

$$G^2(x,1/2) - a(1-x) = F^2(x,1/2) - a(1-x) > 0,$$

we obtain that $G(x,y)$ exists for $y \in [1/2, y_1(x))$,

$$y_1(x) := 1/2 + \big(2\sqrt{a(1-x)}\big)^{-1} \log\frac{F(x,1/2) + \sqrt{a(1-x)}}{F(x,1/2) - \sqrt{a(1-x)}}, \tag{16}$$

and it is given by

$$G(x,y)=\sqrt{a(1-x)}\,\frac{1+\exp(\sqrt{a(1-x)}(2y-1))\frac{F(x,1/2)-\sqrt{a(1-x)}}{F(x,1/2)+\sqrt{a(1-x)}}}{1-\exp(\sqrt{a(1-x)}(2y-1))\frac{F(x,1/2)-\sqrt{a(1-x)}}{F(x,1/2)+\sqrt{a(1-x)}}}. \tag{17}$$

(So $G(x,y)$ blows up as $y\uparrow y_1(x)$.) Consequently $F(x,y)$ exists for $y<y_1(x)$, and $F(x,y)\le G(x,y)$ for $y\in[1/2,y_1(x))$.

**(ii)** Armed with (16)-(17) and $F(x,y)\le G(x,y)$, we choose $x=e^{-n^{-\delta}}$ and $y=e^{0.04an^{-\delta}}$, which is strictly below $y_1(x)$ for $n$ large, as $F(x,1/2)\le 2$, and apply the Chernoff-type bound:

$$\begin{aligned}\mathrm{P}(\mathcal{V}_{n,k}<0.03an)&\le x^{-0.03an}y^{-n}F(x,y)\le x^{-an}y^{-n}G(x,y)\\&=O\big(x^{-an}y^{-n}\big)\le e^{-0.01an^{1-\delta}}.\end{aligned}$$

□

## 3. Closer look at the distribution $\{c_k\}$.

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

**Theorem 3.1.** *The inequality*

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

*holds.*

*Proof.* **(i)** For $n\ge1$, $k\ge0$, 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 $a_{n,k}/n!=E_{n,k}$, 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 $A_k(x)=\sum_{n>0}x^na_{n,k}/n!$ and $B_k(x)=\sum_{n>0}x^nb_{n,k}/n!$; in particular, $B_0(x)=x$. From Lemma 3.1, Lemma 3.2 (Bóna [2]),

$$\begin{aligned}A_k'(x)&=\frac{2}{1-x}\cdot A_k(x)+B_k'(x),\quad(k\ge0),\\B_k'(x)&=2B_{k-1}(x)\cdot\left(\frac{1}{1-x}-\sum_{j=0}^{k-2}B_j(x)\right)-B_{k-1}(x)^2,\quad(k>0).\end{aligned} \tag{18}$$

Introduce $A_{\le k}(x) = \sum_{0\le j\le k} A_j(x)$ and $B_{\le k}(x) = \sum_{0\le j\le k} B_j(x)$. It follows from the equation (18) that

$$A'_{\le k}(x) = \frac{2}{1-x}\cdot A_{\le k}(x) + B'_{\le k}(x), \quad (k \ge 0), \tag{19}$$

$$\frac{d}{dx}\left(\frac{1}{1-x} - B_{\le k}(x)\right) = \left(\frac{1}{1-x} - B_{\le k-1}(x)\right)^2 - 1, \quad (k > 0). \tag{20}$$

The equation (20) can also be obtained directly via an independence argument. Here is how. Let $p_{n,\le k} := \sum_{j\le k} b_{n,j}/n!$ be the probability that the root rank $k$ at most; so $p_{n,>k} := 1 - p_{n,\le k}$ is the probability that the root rank strictly exceeds $k$. So $B_{\le k}(x)$ is the generating function of $\{p_{n,\le k}\}_{n\ge 1}$. Then, for $n > 1$ and $k \ge 0$,

$$p_{n,>k} = (1/n)\sum_{j=0}^{n-1} p_{j,>k-1}\cdot p_{n-j-1,>k-1},$$

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

$$\begin{aligned}\frac{d}{dx}\sum_{n\ge 1} p_{n,>k}x^n &= \left(\sum_{n\ge 0} p_{n,>k-1}x^n\right)^2 - p_{0,>k-1}p_{0,>k-1}\\ &= \left(\sum_{n\ge 0} p_{n,>k-1}x^n\right)^2 - 1.\end{aligned}$$

Here

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

and

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

So

$$\frac{d}{dx}\left(\frac{1}{1-x} - B_{\le k}(x)\right) = \left(\frac{1}{1-x} - B_{\le k-1}(x)\right)^2 - 1.$$

**(ii)** From (19), it follows that, for $k > 0$,

$$A_{\le k}(x) = \frac{1}{(1-x)^2}\int_0^x (1-y)^2 B'_{\le k}(y)\,dy \tag{21}$$
$$= \frac{1}{(1-x)^2}\left[(1-x)^2 B_{\le k}(x) + 2\int_0^x (1-y)B_{\le k}(y)\,dy\right],$$

so for $x \uparrow 1$

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

We conclude that

$$\sum_{j=0}^{k} c_j = 2\int_0^1 (1-y)B_{\le k}(y)\,dy. \tag{22}$$

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

$$1 - \sum_{j=0}^{k} c_j = 2\int_0^1 (1-y)B_{>k}(y)\,dy. \tag{23}$$

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 $k$. A success of this approach depends on how efficient would be our search for a path that has a good chance to be comparable in length to the shortest path.

**(ii)** Here is a randomized greedy algorithm with a plausibly good chance to find such a competitive path. If there are two non-empty 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 $n > 1$, $k \ge 0$,

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

or

$$(n)_2\pi_{n,>k} = 2(n-1)\pi_{n,>k-1} + 2\sum_{j=1}^{n-2}(n-1-j)\pi_{j,>k-1}. \tag{24}$$

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

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

Obviously $B_{>k}(x) \le P_{>k}(x)$, and so the equation (23) yields

$$1 - \sum_{j=0}^{k} c_j \le 2\int_0^1 (1-y)P_{>k}(y)\,dy, \quad (k \ge 0). \tag{25}$$

Since $\pi_{n,>k} = 0$ for $n \le k$, we have $P^{(t)}_{>k}(0) = 0$ for $t \le k$. It follows from (24) that

$$\begin{aligned}
\frac{d^2P_{>k}(x)}{dx^2} &= \sum_{n\ge 2}(n)_2\pi_{n,>k} \\
&= 2\sum_{n\ge 2}(n-1)\pi_{n-1,>k-1}x^{n-2} + 2\sum_{n\ge 2}x^{n-2}\sum_{j=1}^{n-2}(n-1-j)\pi_{j,>k-1} \\
&= 2\,\frac{d}{dx}\sum_{\nu\ge 1}\pi_{\nu,>k-1}x^{\nu} + 2\sum_{j\ge 1}\pi_{j,>k-1}x^j \sum_{n\ge j+2}(n-1-j)x^{n-2-j} \\
&= 2\,\frac{dP_{>k-1}}{dx} + 2\left(\sum_{j\ge 1}\pi_{j,>k-1}x^j\right)\left(\sum_{\nu\ge 1}\nu x^{\nu-1}\right) \\
&= 2\,\frac{dP_{>k-1}}{dx} + \frac{2}{(1-x)^2}\,P_{>k-1}(x).
\end{aligned}$$

Thus

$$\frac{d^2P_{>k}(x)}{dx^2} = 2\,\frac{dP_{>k-1}}{dx} + \frac{2}{(1-x)^2}\,P_{>k-1}(x), \quad (P^{(r)}_{>k}(0) = 0 \text{ for } r \le k). \tag{26}$$

In light of (25) it seems necessary, as before, to integrate successively the differential equations (26) for $P_{>k'}(x)$, $k' = 1, 2, \ldots, k$, and then to evaluate the RHS of the bound (25). In fact, that is how we computed the bounds (25) for $k$ up to 10; linearity of (26) was critical for success of this computation. The data showed, rather compellingly, that the bound decays faster than $(1/2)^k$. In absence of any tractable expression for $P_{>k}(x)$ when $k$ is large, the issue was to find a way to bound the integral in (25) without such an expression.

**(iii)** Linearity of (26) to the rescue again! Introduce

$$I_{k,t} = \int_0^1 (1-y)^t P_{>k}(y)\,dy, \quad k \geq -1,\, t > 0;$$

so

$$1 - \sum_{j=0}^{k-1} c_k \leq 2I_{k,1}, \quad (k > 0). \tag{27}$$

Notice first that, for $t > 0$,

$$\begin{aligned} I_{-1,t} &= \int_0^1 (1-y)^t P_{>-1}(y)\,dy = \int_0^1 \left[-(1-y)^t + (1-y)^{t-1}\right] dt \\ &= \frac{1}{t(t+1)}. \end{aligned} \tag{28}$$

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

$$I_{k,t} = \frac{2}{(t+2)_2}\left[I_{k-1,t} + (t+2)I_{k-1,t+1}\right]. \tag{29}$$

Indeed, using $P_{>k}^{(r)}(0) = 0$ for $r = 0, 1$ and (26),

$$\begin{aligned} I_{k,t} &= \int_0^1 (1-y)^t P_{>k}(y)\,dy \\ &= \frac{1}{(t+2)_2}\int_0^1 (1-y)^{t+2}\frac{d^2 P_{>k}(y)}{dy^2}\,dy \\ &= \frac{2}{(t+2)_2}\int_0^1 (1-y)^{t+2}\left[\frac{dP_{>k-1}(y)}{dy} + \frac{P_{>k-1}(y)}{(1-y)^2}\right] dy \\ &= \frac{2}{(t+2)_2}\left[(t+2)\int_0^1 (1-y)^{t+1} P_{>k-1}(y)\,dy + \int_0^1 (1-y)^t P_{>k-1}(y)\,dy\right] \\ &= \frac{2}{(t+2)_2}\left[I_{k-1,t} + (t+2)I_{k-1,t+1}\right]. \end{aligned}$$

In particular,

$$I_{k,1} = \frac{1}{3}\left[I_{k-1,1} + 3I_{k-1,2}\right] \geq \frac{1}{3}I_{k-1,1},$$

so that $I_{k,1} \geq$ const $(1/3)^k$. We are about to show that in fact $I_{k,1} \leq$ const $(1/3)^k$, i. e. $I_{k,1}$ is of order $(1/3)^k$ exactly.

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

$$I_{k,t} \leq \frac{1}{(t+1)_2}\left(\frac{2}{\tau+2}\right)^{k+1}. \tag{30}$$

By (28), the bound holds for $k = -1$. Inductively, if it holds for for some $k \geq 0$, then by (29)

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

So the bound (30) is proven. In particular,

$$\text{for } t \geq 4, \quad I_{k,t} \leq \frac{1}{(t+1)_2}\left(\frac{1}{3}\right)^{k+1} \Longrightarrow I_{k,4} \leq 0.05\left(\frac{1}{3}\right)^{k+1}.$$

Using the equation (29) for $t = 3$, we have then

$$I_{k,3} = \frac{1}{10}I_{k-1,3} + \frac{1}{2}I_{k-1,4} \leq 0.1I_{k-1,3} + 0.025\left(\frac{1}{3}\right)^{k}.$$

Iterating this recurrence inequality and using (28) for $I_{0,3}$, we obtain

$$\begin{aligned} I_{k,3} &\leq \frac{1}{3\cdot 4}\left(\frac{1}{10}\right)^{k+1} + 0.025\left(\frac{1}{3}\right)^{k}\sum_{j\geq 0}\left(\frac{3}{10}\right)^{j} \\ &= \left(\frac{1}{3}\right)^{k+1}\left(\frac{1}{12} + 0.025\cdot\frac{30}{7}\right) \leq \frac{1}{5}\left(\frac{1}{3}\right)^{k+1}. \end{aligned} \tag{31}$$

Analogously, using the equation (29) for $t = 2$ in conjunction with (31), we iterate the resulting recurrence inequality

$$I_{k,2} \leq \frac{1}{6}I_{k-1,2} + \frac{2}{15}\left(\frac{1}{3}\right)^{k}.$$

Recalling (28) for $I_{0,2}$, we obtain

$$I_{k,2} \leq \left(\frac{1}{3}\right)^{k+1}. \tag{32}$$

Lastly, combining (29) for $t = 1$ and (32), we have

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

this recurrence, and using (28) for $I_{0,1}$, we arrive at

$$I_{k,1} \leq \frac{6k+7}{6}\left(\frac{1}{3}\right)^{k}. \tag{33}$$

The bounds (33) and (27) taken together imply

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

□

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 $\alpha_0$ denote the smaller root; $\alpha_0 \approx 0.373$.

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

$$1 - \sum_{j=0}^{k} c_j \ge \gamma e^{-k/\alpha_0}. \tag{34}$$

*Proof.* **(a)** 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 $s \in [0, m-1]$,

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

where $X_{m,\mu}$ is the total number of leaves at distance $\mu$ from the root. By (39),

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

Given $\gamma > 0$, we have: for $\mu \le \gamma \log m$,

$$\begin{aligned}\mathrm{E}[X_{m,\mu}] &\le \frac{\mu}{m(\log m + 1)} \cdot \frac{2^\mu(\log m+1)^\mu}{\mu!} \\ &\le \frac{\gamma e^\gamma}{m}\left(\frac{2\log m}{\mu/e}\right)^\mu.\end{aligned}$$

As a function of $\mu$, the RHS increases for $\mu \le 2\log m$. Assuming that $\gamma \le 2$, we obtain then that

$$\begin{aligned}\mathrm{P}(S_m \le \gamma \log m) &\le \frac{\gamma^2 e^\gamma \log m}{m} \cdot \left(\frac{2e}{\gamma}\right)^{\gamma \log m} \\ &= \gamma^2 e^\gamma \log m \cdot \exp[g(\gamma)\log m].\end{aligned}$$

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

$$\mathrm{P}(S_m \le (\alpha_0/2)\log m) \le (\alpha_0/2)^2 e^{\alpha_0/2} \cdot m^{g(\alpha_0/2)} \log m = o(1). \tag{35}$$

For $\alpha := \mu/\log m \in (\alpha_0/2, \alpha_0)$ we have (see [9])

$$\begin{aligned}\mathrm{E}[X_{m,\mu}] &= (1+\varepsilon_m)K(\alpha)(\log m)^{-1/2}\exp[g(\alpha)\log m], \\ K(\alpha) &:= \left(\sqrt{2\pi\alpha}\,\Gamma(\alpha)\right)^{-1}\exp(\alpha - 1),\end{aligned} \tag{36}$$

where $\lim_{m\to\infty}\varepsilon_m = 0$. By 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 $\sum_{\alpha\in(\alpha_0/2,\alpha_0]} \mathrm{E}[X_{m,\mu}]$ is of order

$$(\log m)^{-3/2} \int\limits_{\alpha_0 g'/2}^{\alpha_0 g'} e^{-x}\,dx = O\big((\log m)^{-3/2}\big).$$

Recalling (35), we conclude that

$$\mathrm{P}(S_m \le \alpha_0 \log m) = O\big((\log m)^{-3/2}\big). \tag{37}$$

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

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

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

$$\frac{\mathrm{E}[Y_{n,\ell}]}{n+1} = \frac{2}{\ell+1} - \frac{1}{n+1} \Longrightarrow \frac{\mathrm{E}[Y_{n,\ell}]}{n} = (1+1/n)\left(\frac{2}{\ell+1} - \frac{1}{n+1}\right). \tag{38}$$

Consider a generic subtree on $m \ge \ell$ vertices. Conditioned on its vertex set $\{p(i_1),\dots,p(i_m)\}$, $i_1 < \dots < i_m$, this subtree has the same distribution as the tree on $[m]$ grown from the 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 (37), we have: uniformly for $m \ge \ell$,

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

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

$$\sum_{j\ge \alpha_0 \log \ell} E_{n,k} \ge \mathrm{E}[Z_{n,\ell}].$$

From the above equation it follows that

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

Combining this with (38), we obtain

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

Therefore

$$\begin{aligned}\sum_{j\ge\alpha_0\log\ell} c_j &= \lim_{n\to\infty} n^{-1} \sum_{j\ge\alpha_0\log\ell} E_{n,j}\\ &\ge \liminf_{n\to\infty} \frac{\mathrm{E}[Z_{n,\ell}]}{n}\\ &= \frac{2}{\ell+1}\big[1+O\big((\log\ell)^{-3/2}\big)\big].\end{aligned}$$

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

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

□

By Theorem 3.1 and Theorem 3.2 the radius of convergence of $\sum_k c_k x^k$ is in the interval $[3, e^{1/0.373\ldots}]$. Could the radius actually be equal to $e^{1/0.373\ldots}$?

## 4. Variations

Besides $E_{n,k}$, 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

$$f_k=\lim_{n\to\infty} F_{n,k}/n, \quad g_k=\lim_{n\to\infty} G_{n,k}/n.$$

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 $\{\text{root rank}=k\}$, we have

$$F_{n,k}=f_{n,k}+(1/n)\sum_{j=0}^{n-1}(F_{j,k}+F_{n-1-j,k}), \quad n>1.$$

For $n>0$, $f_{n,0}=0$; for $k>0$, using (7),

$$\begin{aligned} f_{n,k} &\le (n-1)\mathrm{P}(\text{root rank}=k) \le n\mathrm{E}[X_{n,k}] \\ &\le n2^k[x^n]\frac{1}{k!}\left(\log\frac{1}{1-x}\right)^k = 2^k\frac{n}{n!}\,[y^{k-1}](y+1)\cdots(y+n-1) \\ &= 2^k\frac{n(n-1)!}{n!}\left(\sum_{0<i_1<\cdots<i_{k-1}<n}\frac{1}{i_1\cdots i_{k-1}}\right) \qquad (39) \\ &\le \frac{2^k}{(k-1)!}\left(\sum_{1\le i\le n-1}\frac{1}{i}\right)^{k-1} \le \frac{2^k}{(k-1)!}\,(\log n+1)^{k-1} \\ &= O\big((\log n)^{k-1}\big). \end{aligned}$$

So $x_n:=F_{n,k}$, $y_n:=f_{n,k}$ meet the conditions of Lemma 2.2 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 $f_k$, $g_k$, we need the recurrences similar to (19)-(20). Introduce $f_{n,>k}=\sum_{j>k}f_{n,j}$, and $\mathcal{A}_k(x)=\sum_{n\ge1}x^nF_{n,k}$, $\mathcal{B}_k(x)=\sum_{n\ge1}x^nf_{n,k}$, and $\mathcal{B}_{>k}(x)=\sum_{n\ge1}x^nf_{n,>k}$, where $f_{n,>k}:=\sum_{j>k}f_{n,j}$; so

$$\mathcal{B}_k(x)=\mathcal{B}_{>k-1}(x)-\mathcal{B}_{>k}(x). \qquad (40)$$

**Lemma 4.1.**

$$\frac{d}{dx}\mathcal{A}_k(x)=\frac{2}{1-x}\,\mathcal{A}_k(x)+\frac{d}{dx}\mathcal{B}_k(x),\quad (k\ge0), \qquad (41)$$

$$\frac{d}{dx}\mathcal{B}_{>k}(x)=2\left(\frac{1}{1-x}-B_{\le k-1}(x)\right)\mathcal{B}_{>k-1}(x),\quad (k\ge0); \qquad (42)$$

*here $\{B_{\le t}(x)\}$ is the sequence determined by the recurrence* (20)*, $B_{\le-1}(x):=0$, and $\mathcal{B}_{>-1}(x)=\mathcal{B}_{\ge0}(x)$ is the generating function of the expected numbers of leaves, i. e.*

$$\mathcal{B}_{>-1}(x)=\frac{x-1}{3}+\frac{1}{3(1-x)^2}.$$

*Consequently*

$$f_k=2\int_0^1(1-x)\mathcal{B}_k(x)\,dx. \qquad (43)$$

Next, introduce $\widehat{A}_k(x)=\sum_{n\ge1}x^nG_{n,k}$, $\widehat{B}_k(x)=\sum_{n\ge1}x^ng_{n,k}$, where $g_{n,k}:=\mathrm{E}\left[1_{\{R(root)=k\}}\mathcal{L}_n\right]$ and $\mathcal{L}_n$ is the number of leaves closest to the root of the tree.

**Lemma 4.2.**

$$\frac{d}{dx}\widehat{A}_k(x)=\frac{2}{1-x}\,\widehat{A}_k(x)+\frac{d}{dx}\widehat{B}_k(x),\quad (k\ge0), \qquad (44)$$

$$\frac{d}{dx}\widehat{B}_k(x)=2\big[1+B_{\ge k-1}(x)\big]\widehat{B}_{k-1}(x),\quad (k>0,\ \ \widehat{B}_0(x)=x). \qquad (45)$$

*Consequently*

$$g_k = 2\int_0^1 (1-x)\widehat{B}_k(x)\,dx. \tag{46}$$

*Proof.* **(I)** 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 $n \geq 2$, $L_n = L' + L''$ where $L'$ and $L''$ denote the total number of leaves in the left subtree $T'$ and the right subtree $T''$ respectively. Let *root*′ (*root*″ resp.) denote the root of $T'$ ($T''$ resp.) if this subtree is non-empty. If both subtrees are non-empty, then

$$1_{\{R(root>k)\}} = 1_{\{R(root')>k-1\}} \cdot 1_{\{R(root'')>k-1\}}, \quad (k \geq 0).$$

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

$$\begin{aligned}
&\mathrm{E}\big[1_{\{R(root)>k\}}L_n \| J\big] = \mathrm{E}\big[1_{\{R(root')>k-1\}}1_{\{R(root'')>k-1\}}(L'+L'')\|J\big]\\
=&\mathrm{E}\big[1_{\{R((\text{root of } T_j})>k-1\}}1_{\{R((\text{root of } T_{n-1-j})>k-1\}}(L_j + L_{n-1-j})\big]\\
=&\mathrm{E}\big[1_{\{R((\text{root of } T_j})>k-1\}}L_j\big]\mathrm{P}(R((\text{root of } T_{n-1-j}) > k-1))\\
&+ \mathrm{E}\big[1_{\{R((\text{root of } T_{n-1-j})>k-1\}}L_{n-1-j}\big]\mathrm{P}(R((\text{root of } T_j) > k-1))\\
=&f_{j,>k-1}\cdot p_{n-1-j,>k-1} + f_{n-1-j,>k-1}\cdot p_{j,>k-1}.
\end{aligned}$$

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

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

It follows immediately that

$$\frac{d}{dx}\sum_{n\geq 1} f_{n,>k}x^n = 2\left(\sum_{n\geq 0} p_{n,>k-1}x^n\right)\cdot\left(\sum_{n\geq 1} f_{n,>k-1}x^n\right),$$

which is equivalent to (42), since

$$\begin{aligned}
\sum_{n\geq 0} p_{n,>k-1}x^n &= 1 + \sum_{n\geq 1}(1-p_{n,\leq k-1})x^n\\
&= 1 + \frac{x}{1-x} - B_{\leq k-1}(x) = \frac{1}{1-x} - B_{\leq k-1}(x).
\end{aligned}$$

The equation (41) is implied by a simple recurrence

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

Finally, from the equation (41),

$$\begin{aligned} f_k &= \lim_{x\uparrow 1}(1-x)^2\mathcal{A}_k(x) \\ &= \int_0^1 (1-x)^2\frac{d}{dx}\mathcal{B}_k(x)\,dx = 2\int_0^1 (1-x)\mathcal{B}_k(x)\,dx. \end{aligned}$$

The proof of Lemma 4.1 is complete.

**(II)** Let us prove the equation (45). 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 $T'$, empty or not, ($T''$ resp.) closest to its root. Let $k > 0$. If both subtrees are non-empty, i. e. $0 < j < n-1$, then

$$\begin{aligned} &1_{\{R(root=k)\}}\mathcal{L}_n = 1_{\{R(root')=k-1\}}\cdot 1_{\{R(root'')=k-1\}}\mathcal{L}_n \\ &+ 1_{\{R(root')=k-1\}}\cdot 1_{\{R(root'')>k-1\}}\mathcal{L}_n + 1_{\{R(root')>k-1\}}\cdot 1_{\{R(root'')=k-1\}}\mathcal{L}_n \\ =&1_{\{R(root')=k-1\}}\cdot 1_{\{R(root'')=k-1\}}(\mathcal{L}'+\mathcal{L}'') \\ &+ 1_{\{R(root')=k-1\}}\cdot 1_{\{R(root'')>k-1\}}\mathcal{L}' + 1_{\{R(root')>k-1\}}\cdot 1_{\{R(root'')=k-1\}}\mathcal{L}''. \end{aligned}$$

The contribution of the first product on the last RHS to $\mathrm{E}\big[1_{\{R(root)=k\}}\mathcal{L}_n\|J\big]$ is

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

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

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

so that

$$\mathrm{E}\big[1_{\{R(root)=k\}}\mathcal{L}_n\|J\big] = g_{j,k-1}\cdot p_{n-1-j,\geq k-1} + g_{n-1-j,k-1}\cdot p_{j,\geq k-1}.$$

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

$$g_{n,k} = \mathrm{E}\big[1_{\{R(root)=k\}}\mathcal{L}_n\big] = \frac{2}{n}\sum_{j=0}^{n-1} g_{j,k-1}\cdot p_{n-1-j,\geq k-1},$$

and (45) follows immediately. And, as before, the equation (44) is the direct consequence of

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

The equation (46) is proved like the equation (43). This completes the proof of Lemma 4.2. □

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 $\mathcal{L}_{n,k}/V_{n,k}$ 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 4.3.**

$$\lim_{n\to\infty} E\left[\frac{\mathcal{L}_{n,k}}{V_{n,k}}\right] = \frac{f_k}{c_k}, \quad \lim_{n\to\infty} E\left[\frac{\widehat{L}_{n,k}}{V_{n,k}}\right] = \frac{g_k}{c_k}.$$

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

$$\begin{aligned}\mathrm{E}\left[\frac{\mathcal{L}_{n,k}}{V_{n,k}}\right] &= \mathrm{E}\left[\frac{\mathcal{L}_{n,k}}{V_{n,k}} 1_{\{V_{n,k}<0.03an\}}\right] + \mathrm{E}\left[\frac{\mathcal{L}_{n,k}}{V_{n,k}} 1_{\{V_{n,k}\ge 0.03an\}} 1_{\{|V_{n,k}/n - c_k|>\varepsilon\}}\right] \\ &\quad + \mathrm{E}\left[\frac{\mathcal{L}_{n,k}}{V_{n,k}} 1_{\{V_{n,k}\ge 0.03an\}} 1_{\{|V_{n,k}/n - c_k|\le\varepsilon\}}\right] \\ &= E_1 + E_2 + E_3.\end{aligned}$$

Here, by Lemma 2.8 and Corollary 2.7 respectively,

$$E_1 \le n e^{-0.01an^{1-\delta}} \to 0, \quad E_2 = O\big(\mathrm{P}(|V_{n,k}/n - c_k| > \varepsilon)\big) \to 0,$$

as $n \to \infty$, and

$$\begin{aligned}E_3 &= \frac{1}{n(c_k + O(\varepsilon))} \mathrm{E}\left[\mathcal{L}_{n,k} 1_{\{V_{n,k}\ge 0.03an\}} 1_{\{|V_{n,k}/n - c_k|\le\varepsilon\}}\right] \\ &= \frac{\mathrm{E}[\mathcal{L}_{n,k}/n]}{c_k + O(\varepsilon)} \left[1 + O\big(\mathrm{P}(V_{n,k} < 0.03an) + \mathrm{P}(|V_{n,k}/n - c_k| > \varepsilon)\big)\right].\end{aligned}$$

Therefore

$$\lim_{\varepsilon\downarrow 0} \limsup_{n\to\infty} E_3 = \lim_{\varepsilon\downarrow 0} \liminf_{n\to\infty} E_3 = f_k/c_k.$$

So $\lim_{n\to\infty} \mathrm{E}[\mathcal{L}_{n,k}/V_{n,k}] = f_k/c_k$. □

**Note.** A slight modification of the proof of Corollary 2.7 shows that, in probability, $\mathcal{L}_{n,k}/n \to f_k$ and $\widehat{L}_{n,k}/n \to g_k$. Therefore $\mathrm{E}\big[\mathcal{L}_{n,k}/V_{n,k}\big] \to f_k/c_k$ and $\mathrm{E}\big[\widehat{L}_{n,k}/V_{n,k}\big] \to g_k/c_k$ in probability as well.

Using Maple to integrate the differential equations (42) and (45), we computed $\{f_j\}_{j\le 2}$ and $\{g_j\}_{j\le 2}$ via (43) and (46) respectively:

$$\begin{aligned} f_0 &= \frac{1}{3}, \quad f_1 = \frac{17}{30}, \quad f_2 = \frac{152389}{170100}; \\ g_0 &= \frac{1}{3}, \quad g_1 = \frac{1}{3}, \quad g_2 = \frac{49}{180}. \end{aligned} \tag{47}$$

Therefore

$$\begin{aligned} \frac{f_0}{c_0} &= 1, \quad \frac{f_1}{c_1} = \frac{17}{9}, \quad \frac{f_2}{c_2} = \frac{152389}{36141}; \\ \frac{g_0}{c_0} &= 1, \quad \frac{g_1}{c_1} = \frac{10}{9}, \quad \frac{g_2}{c_2} = \frac{2205}{1721}. \end{aligned} \tag{48}$$

**Remarks. (i)** That $g_0$, $g_1$ should both be $1/3$ follows from the observation that, for $n \geq 2$, the number of pairs $(v, u)$, where $v$ is a rank $k$ vertex and $u$ is its closest descendant-leaf, is the same 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

$$\sum_{j=0}^{k} c_j = 2\int_0^1 (1-y)B_{\leq k}(y)\,dy. \tag{49}$$

Then, using using (20) ,

$$\begin{aligned}\int_0^1 (1-y)B_{\leq k}(y)\,dy &= \frac{1}{2}\int_0^1 (1-y)^2 B_{\leq k}(y)'\,dy \\ &= \frac{1}{2}\int_0^1 \left[2-2y+y^2-\big(1-(1-y)B_{\leq k-1}(y)\big)^2\right].\end{aligned}$$

so

$$\sum_{j=0}^{k} c_j = \int_0^1 \left[2-2y+y^2-\big[1-(1-y)B_{\leq k-1}(y)\big]^2\right] dy. \tag{50}$$

The equation (50) allows to compute $c_k$ directly through $B_{\leq k-1}(x)$, 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 $c_4$ equals

$$\frac{1220584641416536621962901132326463044129999022835124255801567873233533770250224491998529007256709600674182808037972317882880000000000},$$

a fraction whose denominator has 67 digits, and whose approximate value is 0.0364. Combining this with the values of $c_i$ for $i \leq 4$, we see that (with high probability) about 99.14 percent of all vertices are of rank four or less.

The prime factorization of the denominator $denom(c_4)$, when $c_4$ is written in simplest terms, obtained by Maple, is even more interesting, since its factorized representation is

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

So the largest prime divisor of $denom(c_4)$ is 31, which is a tiny number compared to $denom(c_4)$. In stark contrast, the numerator of $c_4$ is the product of just two primes, the smaller of which is 232196467. Even more striking is the fact that $denom(c_4)$ is divisible by *every prime* up to 31.

This surprising result warrants a second look at the numbers $c_k$ for $k \leq 3$, already computed in Bóna [2]. Here is the factorized representation of the denominators, including $denom(c_4)$:

- $denom(c_0) = 3$,

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

So all $denom(c_k)$ 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 $denom(c_k)$ 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 got 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 five or less. The number $denom(c_5)$ has 274 digits, and its prime factorization is

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

If not for this strikingly simple factorization, we would not dare to type in the 274-long monster. So yet again, $denom(c_k)$ 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-based factorization algorithm failed the task.)

Based on these data points, we formulate the following conjectures.

**Conjecture 5.1.** *Let $denom(c_k)$ be the denominator of $c_k$ when $c_k$ is written in smallest terms. Then the largest prime divisor of the denominator is at most as large as some relatively slowly growing function of $k$, possibly $2^{k+1} + 1$.*

**Conjecture 5.2.** *Let $k \geq 2$, and let $p_k$ be largest prime divisor of $denom(c_k)$. Then $denom(c_k)$ is divisible by every prime less than $p_k$.*

We are able to prove Conjecture 5.1 but not Conjecture 5.2. 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 sum will not be divisible by at least as high a power of a given prime $p$ as the denominator of $c_k$ is.

In order to prove Conjecture 5.1, we will need a few simple technical lemmas. 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 5.3.** *For all natural numbers $k$, we have $B_k(x) \in \mathbf{PL}$, meaning that $B_k(x)$ is a bivariate polynomial $P_k(u, v)$, at $u = (1 - x)$, $v = \log 1/(1 - x)$.*

*Proof.* See Lemma 4.1 in [2]. □

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

**Lemma 5.4.** *Let $b$ and $c$ be non-negative integers, and let us write*

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

*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 inital case of $c = 0$ being obvious. Indeed, integration by parts yields

$$\begin{aligned} \int (1-x)^b \log\left(\frac{1}{1-x}\right)^c dx = &- \log\left(\frac{1}{1-x}\right)^c \cdot \frac{(1-x)^{b+1}}{b+1} \\ &+ \int \frac{(1-x)^b}{b+1} \cdot c \log\left(\frac{1}{1-x}\right)^{c-1} dx, \end{aligned} \tag{51}$$

and the proof is complete. □

**Note.** The equation (51) implies

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

so iterating the same operation, we obtain

$$I_{b,c} = \frac{c!}{(b+1)^{c+1}}. \tag{52}$$

**Lemma 5.5.** *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 stong 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 prove them for $B_k$. Formula (18) 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 $\sum_{i=1}^m a_i (1-x)^{b_i} \log\left(\frac{1}{1-x}\right)^{c_i}$, where $b_i \geq -1$ is an integer, while $a_i$ is rational and $c_i$ is a nonnegative integer. Moreover, it follows from (18) and the induction hypothesis that, in the 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 $\sum_{i:b_i=-1} a_i (1-x)^{b_i} \log(\frac{1}{1-x})^{c_i}$ to $B_k(x)$ itself is

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

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 5.4 and by (51), we see that in all the summands neither the exponent of $(1-x)$ nor the exponent of $\log \frac{1}{1-x}$ can exceed $2^{k+1} - 1$, since integration of the terms with $b_i \ge 0$, $c_i \ge 0$ will increase these exponents by at most one. 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. □

*Proof.* (of Conjecture 5.1) Recall that (18) implies that

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

Here

$$B_k(x) = \sum_i a_i (1-x)^{b_i}\left(\log\frac{1}{1-x}\right)^{c_i},$$

$0 \le b_i, c_i \le 2^{k+1} - 1$, and no $a_i$ has denominator with a prime divisor larger than $2^{k+1} - 1$. From (52) it follows then that $c_k$ is the sum of rational fractions, 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 $b_i + 2$. □

**Acknowledgement**

The authors are thankful to Frank Garvan who advised them on numerous occasions on how to convince Maple to carry out a difficult task.

## References

[1] D. Aldous, Asymptotic fringe distributions for general families of random trees, *Ann. Appl. Probab.* **1** (1991), no. 2, pp. 228–266.
[2] M. Bóna, $k$-proteced vertices in binary search trees, *Adv. in Appl. Math.* **53** (2014), 111.
[3] L. Devroye, A note on the height of binary search trees, *J. Assoc. Comput. Mach.* **33** (1986), pp. 489–498.
[4] L. Devroye and S. Janson, Protected nodes and fringe subtrees in some random trees, *Electron. Commun. Probab.* **19** (2014), no. 6, 10 pp.
[5] R. R. Du and H. Prodinger, Notes on protected nodes in digital search trees, *Appl. Math. Lett.* **25** (2012), pp. 1025–1028.
[6] P. Flajolet and R. Sedgewick, Analytic Combinatorics, Cambridge University Press, Cambridge, UK, 2009.
[7] H. Kesten and B. Pittel, A local theorem for the number of nodes, the height and the number of final leaves in a critical branching process tree, *Random. Struct. Algorithms.* **8** (1996), pp. 243–299.
[8] V. F. Kolchin, Moment of degeneration of a branching process and height of a random tree, *Math. Notes Acad. Sci. USSR*, **24** (1978), pp. 954–961.
[9] H. Mahmoud and B. Pittel, *SIAM J. Algebraic Discrete Methods*, **5** (1984), pp. 69-81.
[10] H. Mahmoud and M. Ward, Asymptotic distribution of two-protected nodes in random binary search trees, *Appl. Math. Letters.* **25** (2012), no. 12, pp. 2218–2222.
[11] H. Mahmoud and M. Ward, Asymptotic properties of protected notes in random recursive trees. Preprint, 2013.

[12] B. Pittel, Growing Random Binary Trees, *J. Mathematical Analysis and Its Applications*, **103** (1984), pp. 461-480.
[13] B. Pittel, Note on the heights of random recursive trees and random $m$-ary search trees, *Random Struct. Algorithms.* **5** (1994), pp. 337–347.

Department of Mathematics, University of Florida, 358 Little Hall, PO Box 118105, Gainesville, FL, 32611 – 8105 (USA)
*E-mail address*: bona@ufl.edu

Department of Mathematics, The Ohio State University, 231 West 18-th Avenue, Columbus, Ohio 43210 – 1175 (USA)
*E-mail address*: bgp@math.ohio-state.edu