ON THE NUMBER OF VERTICES OF EACH RANK IN PHYLOGENETIC TREES AND THEIR GENERALIZATIONS

MIKLÓS BÓNA

Abstract. We find surprisingly simple formulas for the limiting probability that the rank of a randomly selected vertex in a randomly selected phylogenetic tree or generalized phylogenetic tree is a given integer.

1. Introduction

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 [8] for a comprehensive treatment of these results. The tree height and width are parameters of global nature, see Kolchin [12], Devroye [5], Mahmoud and Pittel [13], Pittel [16], Kesten and Pittel [11], Pittel [17], 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 [14], [15], Bóna [2], Bóna and Pittel [4], Janson and Holmgren [9, 10] and Devroye and Janson [6]. These articles either focused on unlabeled trees, or trees in which every vertex was labeled.

In this paper, we study another natural class of trees, those in which *only the leaves are labeled*. Some trees of this kind have been studied from different aspects. See [3] for a result of the present author and Philip Flajolet on the subject, or Chapter 5 of [18] for enumerative results for two tree varieties of this class.

First, we will consider *phylogenetic trees*, that is, rooted non-plane trees whose vertices are bijectively labeled with the elements of the set $[n] = \{1, 2, \cdots, n\}$, and in which each non-leaf vertex has exactly two children. See Figure 1 for the set of all three phylogenetic trees on label set [3].

We define the *rank* of a vertex as the distance of that vertex from the its closest descendent leaf, so leaves have rank 0, neighbors of leaves have rank

Date: May 27, 2015.

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.

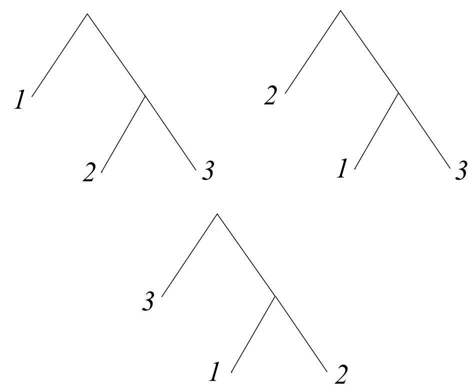

FIGURE 1. The three phylogenetic trees on leaf set [3].

1, and so on. Then for each fixed i, we are able to prove that as n goes to infinity, the probability that a random vertex of a random phylogenetic tree on label set $[n]$ is of rank i converges to a limit $P_{2,i}$, and we are able to compute that limit. The obtained numerical values will be much simpler than the numerical values obtained for other tree varieties, for instance in [2] or [4]. Indeed, we will prove that $P_{2,i} = 2^{1-2^i} - 2^{1-2^{i+1}}$. This will follow from the even simpler formula 2^{1-2^i} for the probability that a random vertex in a random phylogenetic tree is of rank *at least* i.

Then we consider the same questions for *generalized* phylogenetic trees. In such trees, every non-leaf vertex has exactly k children, for some fixed positive integer $k \geq 2$, and the rest of the definition is unchanged. The proofs will be somewhat more involved than those for the special case of $k = 2$, since we will not have explicit, closed forms for our generating functions, essentially because quadratic equations are easier to handle than algebraic equations of a higher degree. Nevertheless, using the Lagrange inversion formula and some observations, we will be able to prove results that are just as precise as those we obtain for the case of $k = 2$.

These results are notable for several reasons. First, the obtained formulas are surprisingly simply. Second, the numbers $P_{k,i}$ decrease very fast, in a doubly exponential way. To compare, note that in [4], the corresponding numbers for binary search trees are shown to decrease in a simply exponential way. Third, the obtained explicit formulas make it routine to prove that the sequence $P_{k,i}$ is log-concave for any fixed i, a fact that is plausible to conjecture, but probably hopeless to prove, for many other tree varieties. Fourth, in the last section we will show an example to illustrate that even for phylogenetic trees, a result that is numerically so simple cannot be expected.

We end the paper by a few open questions asking for combinatorial proofs of some of the mentioned phenomena.

2. Warming up: the case of $k = 2$

In this section, we study the special case of $k = 2$. The results of this section are all special cases of the more general results of the next section. However, our methods in this section are much more explicit, so readers who want to see an example of a more direct case should read this section first, while readers who prefer the "big picture" may skip ahead to the next section, then return to this one for examples.

2.1. **Preliminaries.** Let $T(x)$ be the exponential generating function for the numbers t_n of all phylogenetic trees whose leaves are bijectively labeled with the elements of $[n]$, with $t_0 = 0$. Removing the root of such a tree, we either get the empty set, or a pair of phylogenetic trees so that the union of the set of labels of their leaves is $[n]$. Therefore, the product formula of exponential generating functions (see for instance Chapter 5 of [18]) shows that $T(x) = x + T^2(x)/2$. Solving this quadratic equation for $T(x)$ and using the initial condition $T(0) = 0$, we get that $T(x) = 1 - \sqrt{1-2x}$, so $t_1 = 1$, and $t_n = (2n-3)!! = 1 \cdot 3 \cdot 5 \cdots (2n-3)$ if $n \geq 2$.

As a practical matter, it turns out to be advantageous to count vertices that are *at least* of a given rank. (For most other tree varieties, this is not the case.) Let $m_i(n)$ be the total number of vertices of rank at least i in all phylogenetic trees on leaf set $[n]$. Let $M_i(x) = \sum_{n\geq 1} m_i \frac{x^n}{n!}$. As each such tree has a total of $2n-1$ vertices, it follows that $m_0(n) = (2n-1)!!$ for $n \geq 1$, so $M_0(x) = \sum_{n\geq 1}(2n-1)!!\frac{x^n}{n!} = 1/\sqrt{1-2x} - 1$.

We will need the following straightforward, but important, fact about $M_0(x)$. Let $[x^n]f(x)$ denote the coefficient of x^n in the power series $f(x)$.

Proposition 2.1. *For all positive integers q, the equality*

$$\lim_{n\to\infty} \frac{[x^n]x^q M_0(x)}{[x^n]M_0(x)} = \frac{1}{2^q} \tag{1}$$

holds.

Proof. On the one hand, $[x^n]M_0(x) = \frac{(2n-1)!!}{n!}$. On the other hand,

$$x^q M_0(x) = \sum_{n\geq 1}(2n-1)!!\frac{x^{n+q}}{n!} = \sum_{n\geq 1}(2n-2q-1)!!\frac{x^n}{(n-q)!}.$$

So the fraction on the left-hand side of (1) is equal to

$$\frac{(2n-2q-1)!!}{(n-q)!}\cdot\frac{n!}{(2n-1)!!} = \frac{n(n-1)\cdots(n-q+1)}{(2n-1)(2n-3)\cdots(2n-2q+1)},$$

and our claim is proved. □

In order to compute $M_i(x)$ for larger values of i, we need some more notation. Let $r_i(n)$ be the number of phylogenetic trees on leaf set $[n]$ in which the root is of rank at least i. Let $R_i(x) = \sum_{n\geq 1} r_i \frac{x^n}{n!}$. For instance, $r_0(n) = (2n-3)!! = t_n$ for all n, so $R_0(x) = T(x) = 1 - \sqrt{1-2x}$. Similarly,

$r_1(n) = t_n$ if $n \geq 2$, and $r_1(1) = 0$ (since in the one-vertex tree, the root is of rank zero), leading to the formula $R_1(x) = T(x) - x = 1 - x - \sqrt{1-2x}$.

Lemma 2.2. *For $i \geq 0$, the equality*

$$M_i(x) = M_i(x) \cdot T(x) + R_i(x) \tag{2}$$

holds.

Proof. Note that $m_i(n)$ is the number of ordered pairs (v, T), where T is a phylogenetic tree on leaf set $[n]$, and v is a vertex of rank at least i in T. If v is not the root of T, then removing the root of T, we get two subtrees, one with the marked vertex v, and one without marked vertices. By the product formula of exponential generating functions, such pairs are counted by the generating function $M_i(x)T(x)$, which is the product of the generating functions of the two components. If v is the root of T, then v contributes one to the count of vertices of rank at least i. This happens $r_i(n)$ times, since it happens once for each tree counted by $r_i(n)$. □

Expressing $M_i(x)$ from (2), we get that for all $i \geq 0$, the equality

$$M_i(x) = \frac{R_i(x)}{1 - T(x)} = \frac{R_i(x)}{\sqrt{1-2x}} \tag{3}$$

holds.

For instance,

$$M_1(x) = \frac{R_1(x)}{\sqrt{1-2x}} = \frac{1 - x - \sqrt{1-2x}}{\sqrt{1-2x}} = \frac{1-x}{\sqrt{1-2x}} - 1.$$

From this formula, it is easy to compute that $m_1(n) = (n-1) \cdot (2n-3)!!$. We have of course known this anyway, since each tree counted by $m_1(n)$ has $2n - 1$ vertices, n of which are leaves.

The generating functions $R_i(x)$ are easy to compute for higher i as well.

Proposition 2.3. *For all $i \geq 1$, the recurrence relation*

$$R_i(x) = \frac{R_{i-1}^2(x)}{2} \tag{4}$$

holds.

Proof. The root of a phylogenetic tree is of rank at least i if and only if both children of that root are of rank at least $i - 1$. Our claim is therefore an immediate consequence of the product formula, noting that our trees are non-plane. □

Corollary 2.4. *For all $i \geq 0$, the equality*

$$R_i(x) = \frac{T(x)^{2^i}}{2^{2^i - 1}} = \frac{(1 - \sqrt{1-2x})^{2^i}}{2^{2^i - 1}} \tag{5}$$

holds.

Proof. Straightforward by induction using Proposition 2.3 and the fact that $T(x) = R_0(x)$. □

Corollary 2.5. *For all $i \geq 0$, the equality*

$$M_i(x) = \frac{1}{2^{2^i - 1}} \cdot \frac{(1 - \sqrt{1-2x})^{2^i}}{\sqrt{1-2x}} \tag{6}$$

holds.

Proof. Immediate by comparing the results of Lemma 2.2 and Corollary 2.4. □

Now we have reached the surprising part of this section. In the next lemma, we will prove that the asymptotic behavior of the coefficients of the second term of the right-hand side of (6) is very similar to that of the coefficients of $M_0(x)$. In fact, the limit of the ratios of these coefficient sequences is 1.

Lemma 2.6. *For all $i \geq 1$, we have*

$$\lim_{n\to\infty} \frac{m_i(n)}{m_0(n)} = \lim_{n\to\infty} \frac{m_i(n)}{(2n-1)!!} = \frac{1}{2^{2^i-1}}.$$

That is, for all $i \geq 1$, the probability that a randomly selected vertex of a randomly selected phylogenetic tree is of rank at least i converges to 2^{1-2^i} as n goes to infinity.

Proof. Let $A(x) = \frac{(1-\sqrt{1-2x})^{2^i}}{\sqrt{1-2x}}$. We will split $A(x)$ into the sum of three expressions, and we will show that two of those three components will have a negligible contribution to $[x^n]A(x)$.

Let us expand the numerator of $A(x)$ by the binomial theorem. We get a sum of terms of the form $\binom{2^i}{j}(\sqrt{1-2x})^j$, where j goes from 0 to 2^i. When we divide this sum by the denominator $\sqrt{1-2x}$, the odd powers of $\sqrt{1-2x}$ in the numerator will become even powers of $\sqrt{1-2x}$, in other words, powers of $(1-2x)$, yet in other words, polynomials of degree at most 2^{i-1}. If $n > 2^{i-1}$, then the sum $p_1(x)$ of these polynomials will not contribute to $[x^n]A(x)$.

Therefore, for large enough n, we can determine $[x^n]A(x)$ by determining the coefficients of x^n in the power series

$$\sum_{\substack{j=0 \\ j \text{ even}}}^{2^i} \frac{\binom{2^i}{j}(\sqrt{1-2x})^j}{\sqrt{1-2x}}. \tag{7}$$

Note that by the binomial theorem, the last displayed sum is equal to

$$G_i(x) = \frac{g_i(x)}{\sqrt{1-2x}} = \frac{(1-\sqrt{1-2x})^{2^i} + (1+\sqrt{1-2x})^{2^i}}{2\sqrt{1-2x}}. \tag{8}$$

Now we claim that for all $i \geq 2$, the polynomial

$$g_i(x) - 2^{2^{i-1}} x^{2^{i-1}}$$

is divisible by $(1-2x)$. *In other words, if we remove the term of highest degree from $g_i(x)$, the remaining polynomial is divisible by $1-2x$.* This is somewhat surprising.

In order to see this, note that

$$g_i(x) - 2^{2^{i-1}} x^{2^{i-1}} = (g_i(x) - 1) + \left(1 - 2^{2^{i-1}} x^{2^{i-1}}\right).$$

We show that both summands on the right-hand side are divisible by $1-2x$. Indeed, looking at (8), we see that 1 is the only expansion term in $g_i(x)$ that is not divisible by $1-2x$ that does not cancel. So $g_i(x) - 1$ is divisible by $1-2x$. On the other hand,

$$1 - 2^{2^{i-1}} x^{2^{i-1}} = (1 - 2^{2^{i-2}} x^{2^{i-2}}) \cdot (1 + 2^{2^{i-2}} x^{2^{i-2}}),$$

and our claim is proved by induction on i, the case of $i = 2$ being trivial.

To summarize, we have shown that

$$A(x) = p_1(x) + p_2(x)\sqrt{1-2x} + \frac{2^{2^{i-1}} x^{2^{i-1}}}{\sqrt{1-2x}},$$

where p_1 and p_2 are polynomial functions.

As $p_1(x)$ is a polynomial, it does not contribute to $[x^n]A(x)$ if n is large enough. As $\sqrt{1-2x} = 1 - \sum_{n\geq 1}(2n-3)!!x^n/n!$, the contribution of $p_2(x)\sqrt{1-2x}$ to $[x^n]A(x)$ is negligible compared to the contribution of $1/\sqrt{1-2x} = \sum_{n\geq 0}(2n-1)!!x^n/n!$. On the other hand, setting $q = 2^{i-1}$ and applying Proposition 2.1, the contribution of $\frac{2^{2^{i-1}} x^{2^{i-1}}}{\sqrt{1-2x}}$ to $[x^n]A(x)$ is asymptotically equal to $m_0(n)/n! = (2n-1)!!/n!$. As $M_i(x) = \frac{A(x)}{2^{2^i-1}}$, our claim is proved. □

That is, Lemma 2.6 proves that as n goes to infinity, about 1/8 of all the vertices of all phylogenetic trees on label set $[n]$ will be of rank at least two, about 1/128 of the vertices will be of rank at least three, and so on, and in general, about 2^{1-2^i} of all vertices will be of rank at least i. (We have already known that about half of the vertices will be of rank at least 1.)

Corollary 2.7. *Let $p_{2,i}(n)$ be the probability that a randomly selected vertex of a random phylogenetic tree on label set $[n]$ is of rank exactly i. Then $P_{2,i} = \lim_{n\to\infty} p_{2,i}(n)$ exists, and*

$$P_{2,i} = \lim_{n\to\infty} \frac{m_i(n) - m_{i+1}(n)}{m_0(n)} = \frac{1}{2^{2^i-1}} - \frac{1}{2^{2^{i+1}-1}}.$$

So about half of the vertices are leaves, about 3/8 of the vertices are of rank two, about 7/128 of the vertices are of rank three, and so on.

3. The general case

3.1. Generalized phylogenetric trees and the Lagrange inversion formula. In this section we consider generalized phylogenetic trees. The only change in the definition is that now each non-leaf vertex must have exactly k children, for some fixed integer $k \geq 2$. Let $t_{k,n}$ be the number of such trees on leaf set $[n]$, and set $t_{k,0} = 0$. Let $T_k(x)$ be the exponential generating function of these numbers.

Removing the root of such a tree, we get either the empty set, or an unordered set of k such trees, leading to the functional equation

$$T_k(x) = x + \frac{T_k^k(x)}{k!}. \tag{9}$$

This means that $T_k(x)$ is the compositional inverse of the power series $F_k(x) = x - x^k/k!$, so the coefficients of $T_k(x)$ can be computed by the Lagrange inversion formula. However, that does not imply that the power series $T_k(x)$ has a simple closed form. In fact, it usually does not, since it is a solution of a functional equation of degree k, where k can be arbitrarily high.

We extend our notation from the previous section as follows. Let $m_{i,k}(n)$ denote the number of all vertices that are of rank at least i in all k-phylogenetic trees on leaf set $[n]$. Let $M_{i,k}(x)$ be the exponential generating function of the numbers $m_{i,k}(n)$. Similarly let $r_{i,k}(n)$ be the number of k-phylogenetic trees on leaf set $[n]$ in which the root is of rank at least i, and let $R_{i,k}(x)$ be the exponential generating function of the numbers $r_{i,k}(n)$.

While the Lagrange inversion formula cannot provide a closed form for most of our generating functions, it is still useful for us in that it enables us to prove the following useful proposition. We include the proof of the proposition, but it can be skipped without causing difficulties in reading the rest of the paper.

Proposition 3.1. *Let p be a polynomial function. Then*

$$\lim_{n\to\infty} \frac{[x^n]p\,(T_k(x))}{[x^n]M_{0,k}(x)} = 0.$$

Proof. Note that $[x^n]M_{0,k}(x)$ as well as $[x^n]T_k(x)$, and hence, $[x^n]p\,(T_k(x))$ are nonzero if and only if $n-1$ is divisible by $k-1$. Indeed, growing a k-phylogenetic tree from a single root by turning leafs into parents of leaves, each step will increase the number of leaves by $k-1$.

Clearly, it suffices to prove the statement in the special case when $p(x) = x^\ell$, that is, when $p(T_k(x)) = T_k^\ell(x)$. Indeed, all polynomials are linear combinations of such monomials with constant coefficients. We can also assume that $\ell > 0$, since the stament is obviously true for the polynomial $x^0 = 1$.

We use the following version of the Langrange inversion formula (see Chapter 5 of [18] for a proof. Let n and ℓ be positive integers, and let

$F^{\langle -1\rangle}(x)$ be the compositional inverse of the power series $F(x)$. Then

$$n[x^n](F^{\langle -1\rangle}(x))^\ell = \ell[x^{n-\ell}]\left(\frac{x}{F(x)}\right)^n. \tag{10}$$

Setting $F(x) = F_k(x) = x - \frac{x^k}{k!}$, and recalling that $F^{\langle -1\rangle}(x) = T_k(x)$, formula (10) yields

$$n[x^n]T_k^\ell(x) = \ell[x^{n-\ell}]\left(\frac{x}{x-\frac{x^k}{k!}}\right)^n.$$

From this, we compute

$$\begin{aligned}[x^n]T_k^\ell(x) &= \frac{\ell}{n}[x^{n-\ell}]\left(1-\frac{x^{k-1}}{k!}\right)^{-n}\\ &= \frac{\ell}{n}[x^{n-\ell}]\sum_{s\geq 0}\binom{-n}{s}\left(-\frac{x^{k-1}}{k!}\right)^s\\ &= \frac{\ell}{n}[x^{n-\ell}]\sum_{s\geq 0}\binom{n+s-1}{s}\frac{x^{s(k-1)}}{k!^s}\end{aligned}$$

So, setting $n-\ell = s(k-1)$, we have $n = s(k-1)+\ell$, and the last displayed chain of equalities implies that

$$[x^n]T_k^\ell(x) = \frac{\ell}{s(k-1)+\ell}\binom{ks+\ell-1}{s}\frac{1}{k!^s}. \tag{11}$$

Note that in particular, for $\ell = 1$, we get

$$[x^n]T_k(x) = \frac{1}{s(k-1)+1}\binom{ks}{s}\frac{1}{k!^s}. \tag{12}$$

On the other hand, as $M_{0,k}(x)$ counts all vertices of all k-phylogenetic trees on leaf set $[n]$. As we said at the begining of this proof, this implies that $n = (k-1)s+1$, for some nonnegative integer s, and it is easy to see that such trees have exactly s non-leaf vertices, and therefore, $ks+1$ total vertices. So each coefficient of $M_{0,k}$ is $ks+1$ times as large as the corresponding coefficient of $T_k(x)$.

Therefore, it follows from (12) that

$$[x^n]M_{0,k}(x) = (ks+1)\frac{1}{s(k-1)+1}\binom{ks}{s}\frac{1}{k!^s}.$$

Comparing this with (11), we get that

$$\frac{[x^n]\left(T_k(x)^\ell\right)}{[x^n]M_{0,k}(x)} = \frac{\frac{\ell}{s(k-1)+\ell}\binom{ks+\ell-1}{s}\frac{1}{k!^s}}{(ks+1)\frac{1}{s(k-1)+1}\binom{ks}{s}\frac{1}{k!^s}}$$

$$= \frac{1}{ks+1}\cdot\frac{(s(k-1)+1)\ell}{s(k-1)+\ell}\cdot\frac{(ks+\ell-1)(ks+\ell-2)\cdots(ks+\ell-s)}{(ks)(ks-1)\cdots(ks-s+1)}.$$

As n goes to infinity, so does $n-1=(k-1)s$, and therefore, ks. So the product in the last displayed line clearly converges to 0, since the first term converges to 0, the second one converges to the fixed integer ℓ, and the third one converges to 1. □

3.2. **Generalized versions of basic counting results.** Now we announce the generalized versions of some facts that we have proved in the special case of $k=2$.

Lemma 3.2. *For $i \geq 0$, the equality*

$$M_{i,k}(x) = M_{i,k}(x) \cdot \frac{T_k(x)^{k-1}}{(k-1)!} + R_{i,k}(x)$$

holds.

Proof. Removing the root of a k-phylogenetic tree in which one non-root vertex of rank at least i is marked, we get one such tree with one marked vertex of rank at least i, and an unordered set of $k-1$ trees with no marked vertices. By the product formula of exponential generating functions, such collections have generating function $M_{i,k}(x) \cdot \frac{T_k(x)^{k-1}}{(k-1)!}$. On the other hand, trees in which the root is marked and is of rank at least i are simply counted by $R_{i,k}(x)$. □

Therefore,

$$M_{i,k}(x) = \frac{R_{i,k}(x)}{1 - \frac{T_k(x)^{k-1}}{(k-1)!}}. \tag{13}$$

Proposition 3.3. *For all $i \geq 1$, the recurrence relation*

$$R_{i,k}(x) = \frac{R^k_{i-1,k}(x)}{k!} \tag{14}$$

holds.

Proof. Removing the root of a k-phylogenetic tree in which the root has rank at least i, we get an unordered set of k such trees in which the root has rank at least $i-1$. The claim now follows from the product formula. □

Let us introduce the notation

$$c_i = c_{i,k} = \frac{k^i - 1}{k-1}$$

for shortness.

Corollary 3.4. *For all $i \geq 0$, the equality*

$$R_{i,k}(x) = \frac{T_k(x)^{k^i}}{k!^{c_i}} \tag{15}$$

holds.

Proof. This is straightforward by induction. Indeed, for $i = 0$, the equality $R_{i,k}(x) = T_k(x)$ holds, since in each tree, the root is of rank at least 0. Let us assume that the statement is true for $i-1$, that is,

$$R_{i-1,k}(x) = \frac{T_k(x)^{k^{i-1}}}{k!^{c_{i-1}}}.$$

Now take the kth power of both sides, then divide by $k!$. By Proposition 3.3, this turns the left-hand side into $R_{i,k}(x)$, so we get the equality

$$R_{i,k}(x) = \frac{T_k(x)^{k^i}}{k!^{kc_{i-1}+1}}.$$

This proves our claim since $kc_{i-1} + 1 = c_i$. □

Corollary 3.5. *For all $i \geq 0$, the equality*

$$M_{i,k}(x) = \frac{1}{k!^{c_i}} \cdot \frac{T_k(x)^{k^i}}{1 - \frac{T_k(x)^{k-1}}{(k-1)!}} \tag{16}$$

holds.

In particular, the generating function for the total number of vertices is

$$M_{0,k}(x) = \frac{T_k(x)}{1 - \frac{T_k(x)^{k-1}}{(k-1)!}}. \tag{17}$$

3.3. **Our main results.** Now we are in a position to state and prove the main result of this paper.

Theorem 3.6. *For all integers $k \geq 2$, and for all integers $i \geq 1$, the equality*

$$\lim_{n\to\infty} \frac{m_{i,k(n)}}{m_{0,k}(n)} = \frac{1}{k^{c_i}} = \frac{1}{k^{\frac{k^i-1}{k-1}}} \tag{18}$$

holds.

That is, for large n, about $\frac{1}{k^{c_i}}$ of all vertices are of rank at least i.

Proof. Just like in the special case of $k = 2$, we proceed by splitting a constant multiple of $M_{i,k}(x)$ into two parts, one of which will turn out to be a constant multiple of $M_{0,k}(x)$, and the other one of which will turn out to be negligible, again by a divisibility argument.

To that end, we consider the rightmost factor in (16), and essentially divide the numerator by the denominator, noting that

$$\begin{aligned}
\frac{T_k(x)^{k^i}}{1-\frac{T_k(x)^{k-1}}{(k-1)!}} &= \frac{\left(\frac{T_k(x)^{(k-1)c_i}}{(k-1)!^{c_i}}-1\right)(k-1)!^{c_i}T_k(x)+(k-1)!^{c_i}T_k(x)}{1-\frac{T_k(x)^{k-1}}{(k-1)!}} \\
&= \frac{\left(\frac{T_k(x)^{(k-1)c_i}}{(k-1)!^{c_i}}-1\right)(k-1)!^{c_i}T_k(x)}{1-\frac{T_k(x)^{k-1}}{(k-1)!}} + \frac{(k-1)!^{c_i}T_k(x)}{1-\frac{T_k(x)^{k-1}}{(k-1)!}} \\
&= \frac{\left(\frac{T_k(x)^{(k-1)c_i}}{(k-1)!^{c_i}}-1\right)(k-1)!^{c_i}T_k(x)}{1-\frac{T_k(x)^{k-1}}{(k-1)!}} + (k-1)!^{c_i}M_{0,k}(x).
\end{aligned}$$

We have used (17) in the last step.

Now note that $f^{c_i}-1=(f-1)(f^{c_i-1}+f^{c_i-2}+\cdots+f+1)$. Using this formula for $f=T_k(x)^{k-1}/(k-1)!$, we see that the first summand of the last line in the last displayed array of equations is a *polynomial* function of $T_k(x)$, that is, we have proved that

$$\frac{T_k(x)^{k^i}}{1-\frac{T_k(x)^{k-1}}{(k-1)!}} = p\left(T_k(x)\right)+(k-1)!^{c_i}M_{0,k}(x).$$

By Proposition 3.1, the contribution of $p\left(T_k(x)\right)$ to the coefficient of x^n on the right-hand side is negligible. Comparing this observation with (16) completes the proof. □

Corollary 3.7. *Then for each fixed i, as n goes to infinity, the probability that a random vertex of a random k-phylogenetic tree on label set $[n]$ is of rank i converges to a limit $P_{k,i}$, and*

$$P_{k,i}=\frac{1}{k^{c_i}}-\frac{1}{k^{c_{i+1}}}=\frac{1}{k^{c_i}}-\frac{1}{k^{kc_i+1}}.$$

4. Further directions

The formula $P_{2,i}=2^{1-2^i}-2^{1-2^{i+1}}$ makes it routine to verify the following proposition.

Proposition 4.1. *The sequence infinite sequence $P_{2,i}=2^{1-2^i}-2^{1-2^{i+1}}$ is log-concave as $i=0,1,\cdots$. Similarly, the infinite sequence $P_{k,i}=\frac{1}{k^{c_i}}-\frac{1}{k^{kc_i+1}}$ is log-concave for any fixed k.*

It is plausible to conjecture this log-concave property for the analogously defined sequences for many tree varieties, but it is usually hopeless to prove it. In this case, however, because of the extreme simplicity of the formula for $P_{k,i}$, the log-concave property is routine to prove. Is there a combinatorial proof?

In light of the simple formulas for $P_{2,n}$, one might think that phylogenetic trees are simply too easy, and will lead to simple enumeration formulas for other problems as well. This is not necessarily the case as we will demonstrate. In a phylogenetic tree, let the *height* of a vertex be the length of the *longest* path from a vertex to its closest descendent leaf. Let $L_i(x)$ be the exponential generating function for the total number of vertices of height *at most* i in all phylogenetic trees on leaf set $[n]$. Let $W_i(x)$ be the exponential generating function for the number of phylogenetic trees on leaf set $[n]$ in which the root has height at most i. Then it is straightforward to prove with the methods of Section 2 that for $i \geq 0$, we have

$$L_i(x) = \frac{W_i(x)}{1 - T(x)} = \frac{W_i(x)}{\sqrt{1-2x}}. \tag{19}$$

However, the recurrence relation for the power series $W_i(x)$ is obtained by $W_0(x) = x$, and

$$W_i(x) = \frac{W_{i-1}^2(x)}{2} + x \tag{20}$$

for $i \geq 0$. Indeed, the root of a tree is of height at most i if both of its children are of height at most $i-1$, *or* it does not have any children at all. It is that extra term x on the right-hand side of (20) that makes (20) different from (4), which is the analogous recurrence relation defined for the *rank* of the root. However, this small difference is sufficient to make the formulas for $W_i(x)$ and $L_i(x)$ not as elegant as those for $M_i(x)$.

Indeed, it follows from Proposition 2.1 and identity (19) that the probability that a random vertex of a random phylogenetic tree has height at most i converges to $w_i = W_i(1/2)$. (Note that $W_i(x)$ is a *polynomial* for all i, so $W_i(1/2)$ is defined.) By (20), the sequence of the numbers w_i satisfies $w_0 = 1/2$ and

$$w_i = \frac{w_{i-1}^2 + 1}{2}, \tag{21}$$

and this recurrence relation does not seem to have a closed, explicit solution.

So we have seen that the fact that the formula for $P_{i,2}$, and even for $P_{i,k}$ is so simple is rather exceptional. This raises the question of whether there is a combinatorial proof for this fact that does not use generating functions. Note that a complete proof would also have to show that $P_{i,k}$ exists, not simply compute its numeric value.

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), 1–11.

[3] M. Bóna, P. Flajolet, Isomorphism and symmetries in random phylogenetic trees, *J. Appl. Probab.* **46** (2009), no. 4, 1005–1019.

[4] M. Bóna, B. Pittel, On a random search tree: asymptotic enumeration of vertices by distance from leaves *preprint,* `http://arXiv.org/abs/1412.2796`.

[5] L. Devroye, A note on the height of binary search trees, *J. Assoc. Comput. Mach.* **33** (1986), pp. 489–498.
[6] L. Devroye and S. Janson, Protected nodes and fringe subtrees in some random trees, *Electron. Commun. Probab.* **19** (2014), no. 6, 10 pp.
[7] R. R. Du and H. Prodinger, Notes on protected nodes in digital search trees, *Appl. Math. Lett.* **25** (2012), pp. 1025–1028.
[8] P. Flajolet and R. Sedgewick, Analytic Combinatorics, Cambridge University Press, Cambridge, UK, 2009.
[9] S. Janson, C. Holmgren, Limit Laws for Functions of Fringe trees for Binary Search Trees and Recursive Trees, *preprint*, `http://arXiv.org/abs/1406.6883`.
[10] S. Janson, C. Holmgren, Asymptotic distribution of two-protected nodes in ternary search trees, *preprint* `http://arxiv.org/abs/1403.5573`.
[11] 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.
[12] 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.
[13] H. Mahmoud and B. Pittel, *SIAM J. Algebraic Discrete Methods*, **5** (1984), pp. 69-81.
[14] 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.
[15] H. Mahmoud and M. Ward, Asymptotic properties of protected notes in random recursive trees. Preprint, 2013.
[16] B. Pittel, Growing Random Binary Trees, *J. Mathematical Analysis and Its Applications*, **103** (1984), pp. 461-480.
[17] B. Pittel, Note on the heights of random recursive trees and random m-ary search trees, *Random Struct. Algorithms.* **5** (1994), pp. 337–347.
[18] R. Stanley, Enumerative Combinatorics, Volume 2, Cambridge University Press, Cambridge, UK, 1997.

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