

On the number of vertices of each rank in k -phylogenetic trees

Miklós Bóna

Department of Mathematics, University of Florida, Gainesville, FL, USA

received 10th June 2015, revised 4th Nov. 2015, accepted 17th Mar. 2016.

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

Keywords: search tree, root, leaves, ranks, enumeration, asymptotic, distribution, numerical data

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 [9] for a comprehensive treatment of these results. The tree height and width are parameters of global nature, see [13, 6, 14, 19, 12, 18, 5, 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, [1, 15, 16, 8, 2, 4, 11, 10, 7]. 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 [20] for enumerative results for two tree varieties of this class.

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

We define the *rank* of a vertex as the distance of that vertex from its closest descendent leaf, so leaves have rank 0, neighbors of leaves have rank 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 $P_{k,i}$, 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

$$P_{k,i} = k^{-c_i} - k^{-1-kc_i},$$

where $c_i = c_{i,k} = (k^i - 1)/(k - 1)$. This will follow from an even simpler formula for the probability that a random vertex in a random k -phylogenetic tree is of rank *at least* i .

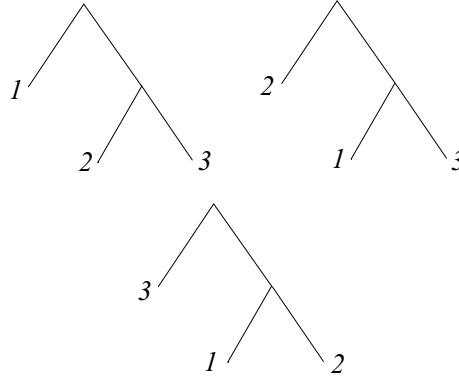


Fig. 1: The three 2-phylogenetic trees on leaf set [3].

The Lagrange inversion formula will be our main tool.

These results are notable for several reasons. First, the obtained formulas are surprisingly simple. 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 2-phylogenetic trees, there are similar questions that lead to much more complicated numerical answers, so the simplicity of our results is surprising.

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

2 Enumeration

2.1 Our trees and the Lagrange inversion formula

Let $t_{k,n}$ be the number of k -phylogenetic trees on leaf set $[n]$, and set $t_{k,0} = 0$. Let $T_k(x) = \sum_{n \geq 0} t_{k,n} \frac{x^n}{n!}$ be the exponential generating function of the sequence 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

$$T_k(x) = x + \frac{T_k^k(x)}{k!}. \quad (1)$$

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.

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 2.1 *Let p be a polynomial function. Then*

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

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 statement is obviously true for the polynomial $x^0 = 1$.

We use the following version of the Lagrange inversion formula (see Chapter 5 of [20] for a proof). Let n and ℓ be positive integers, and let $F^{(-1)}(x)$ be the compositional inverse of the power series $F(x)$. Then

$$n[x^n](F^{(-1)}(x))^\ell = \ell[x^{n-\ell}] \left(\frac{x}{F(x)} \right)^n. \quad (2)$$

Setting $F(x) = F_k(x) = x - \frac{x^k}{k!}$, and recalling that $F^{(-1)}(x) = T_k(x)$, formula (2) 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

$$[x^n]T_k^\ell(x) = \frac{\ell}{s(k-1) + \ell} \binom{ks + \ell - 1}{s} \frac{1}{k!^s}. \quad (3)$$

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

$$[x^n]T_k(x) = \frac{1}{s(k-1) + 1} \binom{ks}{s} \frac{1}{k!^s}. \quad (4)$$

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 beginning 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 (4) 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 (3), we get that

$$\begin{aligned} \frac{[x^n](T_k(x)^\ell)}{[x^n]M_{0,k}(x)} &= \frac{\frac{\ell}{s(k-1)+\ell} \cdot \binom{ks+\ell-1}{s} \cdot \frac{1}{k!^s}}{(ks+1) \cdot \frac{1}{s(k-1)+1} \binom{ks}{s} \cdot \frac{1}{k!^s}} \\ &= \frac{1}{ks+1} \cdot \frac{(s(k-1)+\ell)\ell}{s(k-1)+\ell} \cdot \frac{(ks+\ell-1)(ks+\ell-2)\cdots(ks+\ell-s)}{(ks)(ks-1)\cdots(ks-s+1)}. \end{aligned}$$

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. \square

2.2 Formulas for generating functions

We will now use the tools discussed in Section 2.1 to prove some enumerative lemmas.

Lemma 2.2 *For all integers $k \geq 2$, and for all integers $i \geq 0$, 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 $R_{i,k}(x)$. \square

Therefore,

$$M_{i,k}(x) = \frac{R_{i,k}(x)}{1 - \frac{T_k(x)^{k-1}}{(k-1)!}}. \quad (5)$$

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

$$R_{i,k}(x) = \frac{R_{i-1,k}^k(x)}{k!} \quad (6)$$

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. \square

Let us introduce the notation

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

for shortness.

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

$$R_{i,k}(x) = \frac{T_k(x)^{k^i}}{k!^{c_i}} \quad (7)$$

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 k th power of both sides, then divide by $k!$. By Proposition 2.3, 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 $kc_{i-1} + 1 = c_i$. \square

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

$$M_{i,k}(x) = \frac{1}{k!^{c_i}} \cdot \frac{T_k(x)^{k^i}}{1 - \frac{T_k(x)^{k-1}}{(k-1)!}} \quad (8)$$

holds.

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

$$M_{0,k}(x) = \frac{T_k(x)}{1 - \frac{T_k(x)^{k-1}}{(k-1)!}} \quad (9)$$

2.3 Our main results

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

Theorem 2.6 *For all integers $k \geq 2$, and for all integers $i \geq 1$, the equality*

$$\lim_{n \rightarrow \infty} \frac{m_{i,k}(n)}{m_{0,k}(n)} = \frac{1}{k^{c_i}} = \frac{1}{k^{\frac{k^i-1}{k-1}}} \quad (10)$$

holds.

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

Proof: 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, by a divisibility argument.

To that end, we consider the rightmost factor in (8), 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 (9) in the last step.

Now note that $f^{c_i} - 1 = (f - 1)(f^{c_i-1} + f^{c_i-2} + \dots + f + 1)$. 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

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

By Proposition 2.1, the contribution of $p(T_k(x))$ to the coefficient of x^n on the right-hand side is negligible. Comparing this observation with (8) completes the proof. \square

Corollary 2.7 *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*

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

References

- [1] D. Aldous. Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Probab.*, 1(2):228–266, 1991.
- [2] M. Bóna. k -protected vertices in binary search trees. *Adv. Appl. Math.*, 53:1–11, 2014.
- [3] M. Bóna and P. Flajolet. Isomorphism and symmetries in random phylogenetic trees. *J. Appl. Probab.*, 46(4).
- [4] M. Bóna and B. Pittel. On a random search tree: asymptotic enumeration of vertices by distance from leaves, 2014.

- [5] G.-S. Cheon and L. W. Shapiro. Protected points in ordered trees. *Appl. Math. Lett.*, 21:516–520, 2008.
- [6] L. Devroye. A note on the height of binary search trees. *Assoc. Comput. Mach.*, pages 489–498, 1986.
- [7] L. Devroye and S. Janson. Protected nodes and fringe subtrees in some random trees. *Electron. Commun. Probab.*, 19(6):10 pages, 2014.
- [8] R. R. Du and H. Prodinger. Notes on protected nodes in digital search trees. *Appl. Math. Lett.*, 25:1025–1028, 2012.
- [9] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. CUP, 2009.
- [10] S. Janson and C. Holmgren. Asymptotic distribution of two-protected nodes in ternary search trees. *Electron. J. Probab.*, 20(9):20 pages, 2015.
- [11] S. Janson and C. Holmgren. Limit laws for functions of fringe trees for binary search trees and recursive trees. *Electron. J. Probab.*, 20(4):51 pages, 2015.
- [12] 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*, 8:243–299, 1996.
- [13] V. F. Kolchin. Moment of degeneration of a branching process and height of a random tree. *Math. Notes Acad. Sci. USSR*, 24:954–961, 1978.
- [14] H. Mahmoud and B. Pittel. On the most probable shape of a search tree grown from a random permutation. *SIAM J. Algebraic Discrete Methods*, 5:69–81, 1984.
- [15] H. Mahmoud and M. Ward. Asymptotic distribution of two-protected nodes in random binary search trees. *Appl. Math. Lett.*, 25(12):2218–2222, 2012.
- [16] H. Mahmoud and M. Ward. Asymptotic properties of protected nodes in random recursive trees. *J. Appl. Prob.*, 52(1):290–297, 2015.
- [17] T. Mansour. Protected points in k -ary trees. *Appl. Math. Lett.*, 24(4):478–480, 2011.
- [18] B. Pittel. Growing random binary trees. *J. Mathematical Analysis and Its Applications*, 103:461–480, 1984.
- [19] B. Pittel. Note on the heights of random recursive trees and random m -ary search trees. *Random Struct. Algorithms*, 5:337–347, 1994.
- [20] R. Stanley. *Enumerative Combinatorics*, volume II. CUP, 1997.