Generating Functions for Multi-labeled Trees

É. Czabarka[*]

Department of Mathematics, University of South Carolina, Columbia, SC 29208,
Phone:+1-803-777-7524, FAX: +1-803-777-3783

P.L. Erdős[*]

Rényi Institute of Mathematics, Budapest, Hungary

V. Johnson[*]

Department of Mathematics, University of South Carolina, Columbia, SC 29208

V. Moulton[*]

School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK

[*]Corresponding Author

Email addresses: `czabarka@math.sc.edu` (É. Czabarka), `elp@renyi.hu` (P.L. Erdős), `johnsonv@math.sc.edu` (V. Johnson), `vincent.moulton@cmp.uea.ac.uk` (V. Moulton)

1. Introduction

In evolutionary biology, it is common practice to use *leaf-labeled* (or *phylogenetic*) trees to represent the evolution of species, populations, organisms, and the like [24]. Technically speaking, such a *tree* is a simple, connected graph with no cycles, and it is leaf-labeled in case each of its leaves (i.e. vertices of degree 1) is labeled by precisely one element from some set. The set of labels corresponds to the set of species, populations or organisms under consideration. A simple example of such a tree is presented in Figure 1 (a); we refer reader to [24] for the basic terminology and results on trees and leaf-labeled trees that we shall use in this paper.

Recently it has become apparent that it can also be useful to employ a slightly more general type of tree when trying to understand, for example, gene evolution. In particular, due to processes such as gene (or genome) duplication or lateral gene transfer, trees can often arise in which more than one leaf is labeled by the same element of the label set. We call such trees *leaf-multi-labeled trees* (these are also known as MUL-trees [15]). An example of such a tree, and how it may arise, is presented in Figure 1 (b) and (c). Note that in case each leaf if labeled only once, a leaf-multi-labeled tree is just a leaf-labeled tree. In addition to arising in the study of gene versus species evolution (e.g. [5, 23]), leaf-multi-labeled trees have been used to construct phylogenetic networks (e.g. [13, 15, 18]), and they naturally arise in biogeography (e.g.[8]).

As with leaf-labeled trees, for the purposes of applications it is important to develop a mathematical understanding of leaf-multi-labeled trees. Although at first sight leaf-multi-labeled trees do not seem very different from leaf-labeled trees, the theory of leaf-multi-labeled trees appears to be quite rich in its own right, and several results on theoretical and algorithmic properties of such trees have recently appeared (cf. e.g. [5, 8, 9, 14]). In this paper we will investigate generating functions for such trees, and show how these can be employed to develop recursions for counting them. Note that counting leaf-multi-labeled trees can, for example, be useful for computing bounds on the time required for search algorithms that are commonly used to construct such trees [24].

Counting trees has a rich history. Kirchoff's Laws led to a natural interest in trees and to counting them [17]. At the dawn of graph theory, graphs were considered as 1-dimensional simplicial complexes and of particular interest were the connected ones without cycles, i.e. trees. Various formulae have been developed for counting leaf-labeled trees including the monograph [20]. Cayley [2] formulated that the number of labeled trees on n vertices is n^{n-2}. For example, the well-known formula [3, 6, 7] for the number of binary leaf-labeled trees dates[1] back to Schröder's Fourth Problem [22]. Similar formulae have also been derived for the number of rooted binary leaf-labeled trees [10] (a *rooted tree* is a tree with precisely one distinguished vertex called the *root*).

Concerning generating functions and trees, Harding [10] described ordinary generating functions for rooted, binary *tree-shapes* (i.e. isomorphism classes of unlabeled trees) with or without a specified number of internal vertices. Counting rooted unlabeled trees with the Pólya–Redfield method can be found, e.g., in [19]. Otter's remarkable contribution was counting unrooted unlabeled trees using counting results of rooted unlabeled trees [21]. The functional equation for the ordinary generating function of the number of

---

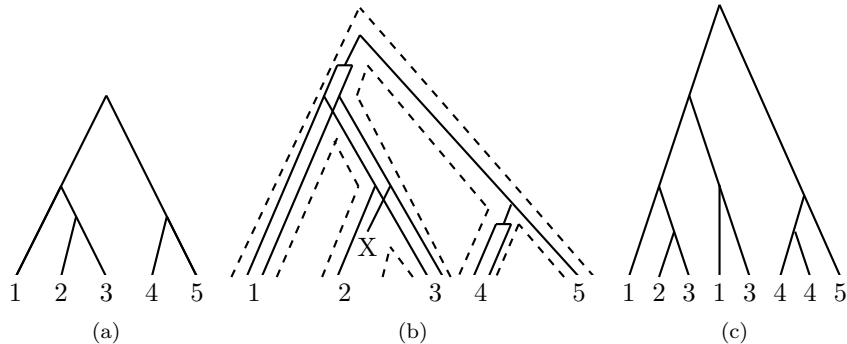[1]A *binary tree* is one in which all non-leaf vertices have degree 3.

Figure 1: (a) A leaf-labeled "species tree" labeled by the set of species $\{1, 2, 3, 4, 5\}$. (b) A "gene tree" (in bold) representing the evolution of a gene, depicted within the species tree (in dashed) from (a) — we see two gene duplication events, and a gene loss (indicated with a cross). (c) The leaf-multi-labeled tree corresponding to the gene tree in (b), for which the label set is $\{1, 2, 3, 4, 5\}$.

rooted unlabeled trees was already known to Cayley (see [21]). Using methods due to Otter and Pólya (described in e.g. [12]), Dobson [4] also gave the generating function for unrooted, binary tree-shapes in terms of Harding's function. In addition, in [24, p.22], a formula involving the exponential generating function for rooted binary trees is given.

Here we shall derive formulae for ordinary generating functions for leaf-multi-labeled trees, and describe how they may be used to develop recursions for counting such trees. As we will only consider ordinary generating functions we shall drop the term "ordinary" from now on; the basics on generating functions that we shall use may be found in, for example, [1].

We now describe the contents of this paper. In Section 2 we give a formula (Theorem 1) involving the generating function for the number of rooted binary leaf-multi-labeled trees, and use this to develop a recursion for counting such trees (see Equation (2)). Note that a tree-shape can be considered as a leaf-multi-labeled tree in which only one label is used to label all leaves, and so Theorem 1 is a direct generalization of Harding's formula for the tree-shape generating function [10] (see also Equation (1)).

In Section 3, we will present a theorem (Theorem 2), which will allow us to relate generating functions of rooted binary leaf-multi-labeled trees with unrooted versions of these trees. This is a generalization of a theorem due to Otter [21] for leaf-labeled trees. Curiously, considering leaf-multi-labeled trees as opposed to leaf-labeled trees allows us to find a simpler proof of Otter's theorem. We then use Theorem 2 in Section 4 to derive a formula for the generating function of unrooted binary leaf-multi-labeled trees (Theorem 5) and some associated recursions (Equation (6)). We conclude by considering non-binary trees, giving formulae for associated generating functions in the rooted (Corollary 1) and unrooted case (Corollary 2). Note that in this section we consider trees where only the root may have degree two, as these are the relevant trees for biological applications. Similar formulae can be developed if we allow internal non-root vertices to have degree two. The interested reader should consult [16].

3

In the rest of this paper, the label set for all of the trees that we shall consider will be $[k] = \{1, 2, \ldots, k\}$, $k \geq 1$. *Semi-labeled trees* are trees where a subset of the leaves have been labeled, and we will allow the labels to repeat. Note that semi-labeled trees may have internal vertices of degree two (and in Theorem 2 this is allowed). Clearly, leaf-multi-labeled trees and unlabeled trees are a subset of semi-labeled trees, where the subset that is labeled is the set of leaves and the empty set, respectively. The notation $\mathrm{Exp}(x)$ will be used for e^x.

2. Rooted binary trees

We begin by considering the generating function for rooted, binary leaf-multi-labeled trees. Note that for technical reasons, we shall consider a single vertex as being a binary, rooted tree. Thus, other than the single vertex, a tree is a rooted (binary) tree if the root vertex has degree at least (equal to) 2, and all the non-root, non-leaf vertices have degree at least (equal to) 3.

Let t_n denote the number of unlabeled rooted binary trees with n leaves (or, equivalently, the number of binary rooted tree-shapes [24]). Note that, as mentioned in the introduction, this corresponds to the number of binary rooted leaf-multi-labeled trees with n leaves on label set [1].

Harding observed [10] (see also Wedderburn [25]) that the generating function for the sequence $\{t_n\}_{n=1}^{\infty}$, viz.

$$T(z) = \sum_{n=1}^{\infty} t_n z^n,$$

satisfies the equation

$$T(z) = z + \frac{1}{2}T^2(z) + \frac{1}{2}T(z^2). \tag{1}$$

A simple justification of this fact is as follows: Clearly, $t_0 = 0$ and $t_1 = 1$. For $n \geq 2$, as the root has degree 2, there are two subtrees below the root and, as the order of the subtrees does not matter, $T^2(z)$ counts all those trees where the two subtrees are not isomorphic twice and those where the subtrees are isomorphic once, whereas $T(z^2)$ counts those trees where the two subtrees are isomorphic.

The same line of reasoning can be used to give a formula for the generating function

$$R(x_1, \ldots, x_k) = \sum_{n_1, \ldots, n_k} r_{n_1, \ldots, n_k} x_1^{n_1} \cdots x_k^{n_k},$$

where $r_{n_1, \ldots, n_k}$ denotes the number of rooted binary leaf-multi-labeled trees on the set $[k]$ with $\sum_{i=1}^{k} n_i$ leaves in which label $j \in [k]$, is used on precisely n_j leaves ($n_j = 0$ is allowed and $r_{0, \ldots, 0} = 0$). In particular, we have

Theorem 1.

$$R(x_1, \ldots, x_k) = (x_1 + \cdots + x_k) + \frac{1}{2}R^2(x_1, \ldots, x_k) + \frac{1}{2}R(x_1^2, \ldots, x_k^2).$$

Although we shall not go into a similar level of detail in the following sections, for the purposes of illustration, we note that this theorem can be used in a straight-forward fashion to obtain a recursion for calculating the numbers $r_{n_1,\dots,n_k}$ as follows. Put

$$h_{n_1,\dots,n_k} = \sum_{m_1=0}^{n_1} \sum_{m_2=0}^{n_2} \cdots \sum_{m_i=0}^{n_i} \cdots \sum_{m_k=0}^{n_k} r_{m_1,\dots,m_k} r_{n_1-m_1,\dots,n_k-m_k}.$$

Then

$$r_{n_1,\dots,n_k} = \begin{cases} 0 & \text{if } \sum_{i=1}^{k} n_i = 0, \\ 1 & \text{if } \sum_{i=1}^{k} n_i = 1, \\ \frac{1}{2}\left(r_{n_1/2,\dots,n_k/2} + h_{n_1,\dots,n_k}\right) & \text{if all } n_i \text{ are even} \\ & \text{and } \sum_{j=1}^{k} n_i \geq 2, \\ \frac{1}{2} h_{n_1,\dots,n_k} & \text{else.} \end{cases} \tag{2}$$

Two special cases of this recursion are worth pointing out. Let $r_{n;k}$ denote the number of rooted binary leaf-multi-labeled trees with n leaves on the set $[k]$, and let $R_k(z)$ be the associated generating function. Then by Theorem 1 we have

$$R_k(z) = kz + \frac{1}{2} R_k^2(z) + \frac{1}{2} R_k(z^2),$$

(just put $x_1 = \dots = x_k = z$) which, in case of $k = 1$ yields (1), as expected. Note that this formula also yields the recursion

$$r_{n;k} = \begin{cases} 0 & \text{if } n = 0, \\ k & \text{if } n = 1, \\ \frac{1}{2} \sum_{j=1}^{n-1} r_{j;k} r_{n-j;k} & \text{if } n > 1 \text{ odd}, \\ \frac{1}{2}\left(r_{n/2;k} + \sum_{j=1}^{n-1} r_{j;k} r_{n-j;k}\right) & \text{else.} \end{cases} \tag{3}$$

As an illustration we present some values of $r_{n;k}$ for $n \leq 10$ and $k \leq 5$ in Table 1. These values were obtained using the program available at the website `http://www.math.sc.edu/~czabarka/programfiles/treecode.html`

Second, we consider the case where we insist on using every label in $[k]$ (i.e. the numbers $r_{n_1,\dots,n_k}$ where each n_i is positive). Then, denoting by $V_k(z)$ the generating function for the binary rooted leaf-multi-labeled trees where the labels come from the set $[k]$ and each label is used, the inclusion-exclusion principle yields

$$V_k(z) = \sum_{n=0}^{\infty} v_{n;k} z^n = \sum_{j=0}^{k-1} (-1)^j \binom{k}{j} R_{k-j}(z).$$

Thus, we obtain the equation

$$v_{n;k} = \sum_{j=0}^{k-1} (-1)^j \binom{k}{j} r_{n;k-j}.$$

5

$n \backslash k$	1	2	3	4	5
1	1	2	3	4	5
2	1	3	6	10	15
3	1	6	18	40	75
4	2	18	75	215	495
5	3	54	333	1260	3600
6	6	183	1620	8010	28275
7	11	636	8202	53240	232500
8	23	2316	43188	366680	1979385
9	46	8610	232947	2590420	17287050
10	98	32763	1282824	18674660	154041450

Table 1: The first few values of $r_{n;k}$, the number of rooted binary leaf-multi-labeled trees with n leaves on the label set $[k]$, obtained using recursion (3).

Again, these numbers can be computed using the program mentioned above.

3. A generalization of Otter's Theorem

In this section we will prove a generalization of Otter's theorem [21, p.589], which is Theorem 2 for unlabeled trees. This will enable us to derive formulae for generating functions associated to non-binary, rooted leaf-multi-labeled trees. We first need to define some additional concepts.

Let $T = (V, E)$ be any (rooted or unrooted) semi-labeled tree. We call the function $\phi : V \to V$ an automorphism of T, if it is a label- and root-preserving graph automorphism, i.e. $xy \in E$ if and only if $\phi(x)\phi(y) \in E$, x and $\phi(x)$ have the same label (or no label), and if there is a root r, then $\phi(r) = r$. We say that $x, y \in V$ are *equivalent* if there is an automorphism ϕ of T such that $\phi(x) = y$; $xy, uv \in E$ are *equivalent* if there is an automorphism ϕ of T such that $\{\phi(x), \phi(y)\} = \{u, v\}$, and $xy \in E$ is a *symmetry-edge* if there is an automorphism ϕ of T such that $\phi(x) = y$ and $\phi(y) = x$.

It is clear that there is at most one symmetry-edge for any tree, as removing a symmetry edge results in a set of two isomorphic trees. It is also obvious that the automorphisms of T form a group, thus the above defined equivalences are equivalence relations on the vertices and on the edges of T. We call the number of equivalence classes the number of *non-isomorphic points* and *non-isomorphic edges*, respectively. The number of non-isomorphic points of T is denoted by p_T, the number of non-isomorphic edges by q_T, and the number of symmetry edges of T by s_T. By the above remarks, $s_T \in \{0, 1\}$.

For an illustration of the idea of equivalence, p_T, q_T and s_T, see the trees depicted on Figure 2.

The following concept will be key in our arguments for relating rooted and unrooted trees. Let T be a tree. A *marking* of T is a choice of one its vertices; the chosen vertex will be called the *marked vertex*. In case T is a rooted tree, we will always assume that the marked vertex is the root.

Now, let T be an (unrooted) binary tree and mark any one of its vertices. Clearly, the number of non-isomorphic markings is p_T. Indeed, marking two different vertices
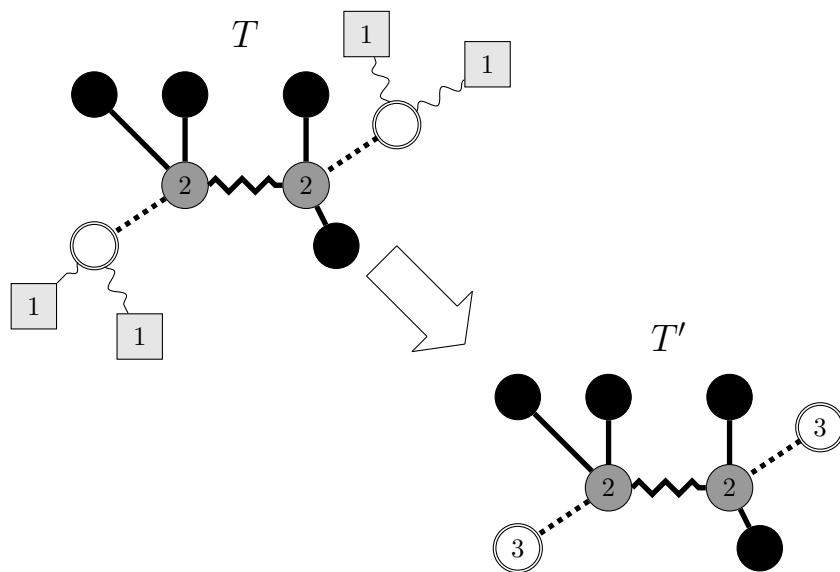
6

Figure 2: A semi-labeled tree T on label set $\{1, 2\}$ and a semi-labeled tree T' on label set $\{1, 2, 3\}$. The shapes, coloring and line types illustrate equivalence: vertices and edges that are depicted by the same kind of shape or line are equivalent. The jagged edge connecting the two vertices labeled by 2 is a symmetry edge. Note that $p_T = q_T = 4$, $s_T = s_{T'} = 1$ and $p_{T'} = q_{T'} = 3$. The parents in T are the white circular nodes connected to the labeled leaves, they form the sets $A = B$ in the proof of Theorem 2. Removing the leaves attached to B and relabeling B as in the proof results in the tree T'.

gives rise to different marked trees precisely when the marked vertices belong to different equivalence classes. Note that by subdividing an edge of T into two edges, and marking the resulting vertex of degree 2, we obtain a rooted binary tree. In particular, q_T corresponds to the number of ways to root the tree T in this way at one of its edges, and s_T corresponds to the number of ways to root the tree T at one of its edges so that the subtrees resulting from removing this root are isomorphic.

With these concepts in hand, we can now prove that Otter's formula (4) for unlabeled trees also holds for semi-labeled trees. It is worthwhile to note that Otter's formula also follows from Theorem 3 at the end of this section.

Theorem 2. *For any semi-labeled tree T we have*

$$p_T - q_T + s_T = 1. \tag{4}$$

Proof. We use induction on the number of vertices n of T.

If $n = 1$, then $p_T = 1$ and $q_T = s_T = 0$. If $n = 2$, then we have two cases. If both the vertices are unlabeled or they have the same label, then $p_T = q_T = s_T = 1$. If only one of the vertices are labeled or they have different labels then $p_T = 2$, $q_T = 1$ and $s_T = 0$. Thus, the theorem is true for $n = 1, 2$.

Let $n \geq 3$ and assume the theorem is true for any semi-labeled tree on less than n vertices. We call a non-leaf vertex a *parent* if at most one of its neighbors are non-leaves.

Note that for any tree on at least 3 vertices the set of parents is nonempty. Indeed, the longest path of the tree has at least 3 vertices; take a vertex that is adjacent to a leaf of a longest path; it is not a leaf, and if it has any neighbors not on the path, that neighbor must be a leaf, otherwise the path could be extended. Therefore the vertex we chose must be a parent.

Also note that, if ϕ is an automorphism of the tree, then ϕ must preserve the degree, the parent property, and the number of leaves that a vertex is adjacent to. Moreover, for any vertex x in the tree, the two multi-sets of the labels of the leaves (where each label appears with the same multiplicity as it is used in the tree) adjacent to x and $\phi(x)$ must be the same. In addition, leaves adjacent to x or $\phi(x)$ are equivalent precisely when they have the same label (or lack-of-label), and leaf-edges adjacent to x or $\phi(x)$ are equivalent precisely when their leaf-endpoints have the same label (or no label).

Let T be a tree on n vertices, and without loss of generality, assume the label set of T is $[k]$. Let A be the set of parents of T. By our above considerations, the set A is nonempty. Fix $x \in A$ and let B be the set of vertices of T that are equivalent to x. Note that $x \in B \subseteq A$. Moreover, if ϕ is an automorphism of T, then ϕ maps B onto B, and if, for some $y \in B$, we have $\phi(y) = z$, then ϕ is a label-preserving bijection from the leaves adjacent to y to the leaves adjacent to z.

Now, define a semi-labeled tree T' as follows: Erase all leaves that are adjacent to vertices of B and label all vertices of B by $k+1$ (a label that was not used before). Note if ϕ is an automorphism of T, then ϕ restricts to an automorphism ϕ' of T'. Moreover, we can extend ϕ' to a label-preserving automorphism of T by taking any vertex $y \in B$ and defining a label-preserving bijection from the leaves adjacent to y to the leaves of $\phi(y)$. (An example for this is provided in Figure 2.)

Since T' has fewer vertices than T, it follows by the induction hypothesis that $p_{T'} - q_{T'} + s_{T'} = 1$. Moreover, in view of our observations above it follows that (i) $p_T = p_{T'} + C$,

and $q_T = q_{T'} + C$, where C is the number of different type of leaves adjacent to elements of B (the type of a leaf is its label if it has one, and "unlabeled" otherwise), and (ii) $s_T = s_{T'}$, and if an edge in T is a symmetry edge, then the same edge is an symmetry edge in T'. Hence $p_T - q_T + s_T = p_{T'} - q_{T'} + s_{T'} = 1$, which completes the proof of the theorem. $\qquad\square$

Note that since unlabeled trees are a subset of semi-labeled trees, Otter's theorem follows from this result. Also note that using a labeling actually makes the proof somewhat easier than the one originally presented by Otter in [21].

It is also worth remarking that it is straight-forward to check that the same proof extends in semi-labeled graphs to give an equation relating the number of dissimilarity points and the number of dissimilarity points within dissimilar 2-connected blocks, as was noted for unlabeled graphs in [11, pp.54-56] (although the proof presented there without using labels is not completely correct, as shown e.g. by a caterpillar tree of diameter 3 on 5 leaves, for details see [16]).

More specifically, let $G = (V, E)$ be a connected semi-labeled graph. A block of G is a maximal 2-connected subgraph of G. We call two points $x, y \in V$ equivalent if there is a label-preserving graph-automorphism ϕ of G such that $\phi(x) = y$. Two blocks, B_1 and B_2 are equivalent if a label-preserving graph-automorphism of G maps the points of B_1 on to B_2. The number of nonequivalent points in G is denoted by p^* and the number of nonequivalent blocks is denoted by b^*. Let $B_1, B_2, \ldots B_{b^*}$ be pairwise nonequivalent blocks in G, and let p_i^* be the number of nonequivalent points in B_i Then it can be shown that

Theorem 3. *For any connected semi-labeled graph G, we have that*

$$p^* - 1 = \sum_{i=1}^{b^*} (p_i^* - 1) \tag{5}$$

In particular, as noted in [11], for a tree T we have $p^* = p_T$, the blocks of a tree are the edges with their endpoints, and thus $b^* = q_T$, and for B_i, we have that $p_i = 2$ if B_i is not the symmetry edge, $p_i = 1$ otherwise. Thus, in this specific case, (5) implies (4).

4. Unrooted binary trees

In this section, we will present formulae involving generating functions for unrooted binary trees.

As indicated in the previous section, to count unrooted binary trees it will be helpful to first count marked trees. Let $m_{n_1,\ldots,n_k}$ be the number of marked, binary leaf-multi-labeled trees where label j is used n_j times, and let $M(x_1, \ldots, x_k) = \sum m_{n_1,\ldots,n_k} x_1^{n_1} \cdots x_k^{n_k}$ to be the corresponding generating function.

Theorem 4.

$$
\begin{aligned}
M(x_1, \ldots, x_k) &= (x_1 + \cdots + x_k)\Big(1 + R(x_1, \ldots, x_k)\Big) + \frac{1}{6} R^3(x_1, \ldots, x_k) \\
&\quad + \frac{1}{2} R(x_1, \ldots, x_k) R(x_1^2, \ldots, x_k^2) + \frac{1}{3} R(x_1^3, \ldots, x_k^3).
\end{aligned}
$$

Proof. Let T be a marked, binary leaf-multi-labeled tree. If the marked vertex x is a leaf of T marked with label j, then either T is a vertex or, by removing x and rooting the resulting tree at its neighbor, we can obtain a rooted binary leaf-multi-labeled tree with the label j used one less time.

It follows that the marked binary leaf-multi-labeled trees where the mark is on a leaf are counted by the generating function $(x_1 + \cdots x_k)(1 + R(x_1, \ldots, x_k))$. So it only remains to describe the generating function for marked trees where an internal vertex (i.e. vertex of degree 3) is marked.

This is determined by the collection of leaf-multi-labeled rooted binary *forests* (i.e. disjoint unions of rooted binary trees) consisting of precisely three rooted binary leaf-multi-labeled trees, since we can obtain such a forest when we remove the marked vertex from T and root the resulting three trees at the neighbor of the marked vertex. Note that, since the neighbor was either a leaf, or it had degree 3, the root so obtained is either a vertex or it has degree 2, as required.

Now, consider the terms $\frac{1}{6}R^3(x_1 \ldots, x_k)$, $\frac{1}{2}R(x_1, \ldots, x_k)R(x_1^2, \ldots, x_k^2)$ and $\frac{1}{3}R(x_1^3, \ldots, x_k^3)$. If all three trees in the forest are non-isomorphic, then the forest is counted by $\frac{1}{6} \cdot 6 = 1$ times by the first term, and is not counted by the rest. If two of the trees in the forest are isomorphic and the third one is not, then the first term counts this forest $\frac{1}{6} \cdot 3 = \frac{1}{2}$ times, the second $\frac{1}{2}$ times and the third term does not count it. And, if all three trees are isomorphic, then the forest is counted $\frac{1}{6} + \frac{1}{2} + \frac{1}{3} = 1$ times by the sum of these three terms. This completes the proof of the theorem. $\square$

Now, denoting by $u_{n_1, \ldots, n_k}$ the number of unrooted leaf-multi-labeled binary trees where the label j is used n_j times, and putting $U(x_1, \ldots, x_k) = \sum u_{n_1, \ldots, n_k} x_1^{n_1} \cdots x_k^{n_k}$, we can use Theorem 2 to obtain the following:

Theorem 5.

$$
\begin{aligned}
U(x_1, \ldots, x_k) &= M(x_1, \ldots, x_k) + (x_1 + \cdots + x_k) - R(x_1, \ldots, x_k) \\
&\quad + R(x_1^2, \ldots, x_k^2) \\
&= \left(R(x_1, \ldots, x_k) + 2 \right) \left(x_1 + \cdots + x_k - 1 + \frac{1}{2} R(x_1^2, \ldots, x_k^2) \right) \\
&\quad + 2 + \frac{1}{3} R(x_1^3, \ldots, x_k^3) + \frac{1}{6} R^3(x_1 \ldots, x_k).
\end{aligned}
$$

Proof. Fix $n_1 \ldots, n_k$ and sum equation (4) over all leaf-multi-labeled binary trees T where for all $j \in [k]$ the label j is used precisely n_j times. If we start from a non-singleton tree, p_T is the number of marked trees that are isomorphic to T, q_T is the number of rooted binary trees that are isomorphic to T after suppressing the root, and s_T is the number of rooted binary trees isomorphic to T, where the two rooted subtrees obtained by removing the root and rooting the remaining trees at the neighbor of the root are isomorphic to one another. So we obtain

$$
u_{n_1, \ldots, n_k} = \begin{cases}
1 & \text{if } \sum n_j = 1, \\
m_{n_1, \ldots, n_k} - r_{n_1, \ldots, n_k} + r_{n_1/2, \ldots, n_k/2} & \text{if } 2 | n_j \text{ for all } j \in [k], \\
m_{n_1, \ldots, n_k} - r_{n_1, \ldots, n_k} & \text{otherwise.}
\end{cases}
$$

We obtain the theorem by multiplying both sides with $x_1^{n_1} \cdots x_k^{n_k}$ and summing over all values of $n_1, \ldots, n_k$. $\square$

10

We note that, letting $u_{n;k}$ be the number of unrooted leaf-multi-labeled binary trees using label set $[k]$ that have n leaves, and putting

$$h^*_{n;k} = kr_{n-1;k} - r_{n;k} + \frac{1}{6} \sum_{i=1}^{n-2} \sum_{j=1}^{n-i-1} \sum_{\ell=1}^{n-i-j} r_{i;k} r_{j;k} r_{\ell;k} + \frac{1}{2} \sum_{\substack{(i,j) \\ 2i+j=n}} r_{i;k} r_{j;k},$$

with $r_{n;k}$ as defined in Section 2, we can use the last theorem to obtain the following recursion for computing $u_{n;k}$.

$$u_{n;k} := \begin{cases} 0 & \text{if } n = 0, \\ k & \text{if } n = 1, \\ h^*_{n;k} + \frac{1}{3} r_{n/3;k} + r_{n/2;k} & \text{if } n = 6\ell, \ \ell \in \mathbb{Z}^+, \\ h^*_{n;k} & \text{if } n = 6\ell \pm 1, \ \ell \in \mathbb{N}, \\ h^*_{n;k} + r_{n/2;k} & \text{if } n = 6\ell \pm 2 \geq 2, \ \ell \in \mathbb{Z}, \\ h^*_{n;k} + \frac{1}{3} r_{n/3;k} & \text{if } n = 6\ell + 3 \geq 2, \ \ell \in \mathbb{Z}. \end{cases} \tag{6}$$

5. Non-binary trees

We have seen how to compute generating functions and recursions for counting rooted and unrooted binary leaf-multi-labeled trees. In this last section, we will consider non-binary trees.

Let $\mathcal{R}$ denote the set of (isomorphism classes) of leaf-multi-labeled rooted trees, which include the single vertex tree and the trees where the degree of every non-root, non-leaf vertex is at least 3, and the root has degree at least 2. Note that for a binary tree with $n \geq 2$ leaves, the number of internal vertices can be given as a function of n ($n-1$ if rooted and $n-2$ if unrooted), but for non-binary trees this is not the case. In particular, an element of $\mathcal{R}$ with $n \geq 2$ leaves can have any number of internal vertices between 1 and $n-1$. Thus it is useful to keep track of the number of internal, unlabeled vertices.

For this reason, we define $a_{u,n_1,\ldots,n_k}$ to be the number of trees in $\mathcal{R}$ with u unlabeled nodes and n_j nodes with label j, and $A(z; x_1, \ldots, x_k)$ be the corresponding ordinary generating function.

We now give a Cayley-type equality for A; the following notation will be helpful. For a leaf-multi-labeled $T \in \mathcal{R}$, we denote by $\ell_j(T)$ the number of vertices that have label j, by $un(T)$ the number of unlabeled vertices, and put

$$\text{term}(T) = z^{un(T)} \prod_{j=1}^{k} x_j^{\ell_j(T)}.$$

Theorem 6.

$$A(z; x_1, \ldots, x_k) = \frac{(x_1 + \cdots + x_k - z) + z \cdot \text{Exp}\left(\sum_{n=1}^{\infty} \frac{1}{n} A(z^n; x_1^n, \ldots, x_k^n) \right)}{z + 1}$$

$$= \left(\sum_{j=0}^{\infty} (-1)^j z^j \right) \left((x_1 + \cdots + x_k - z) + z \cdot \text{Exp}\left(\sum_{n=1}^{\infty} \frac{1}{n} A(z^n; x_1^n, \ldots, x_k^n) \right) \right).$$

11

Proof. There is precisely one tree in $\mathcal{R}$ that is a single vertex and is labeled by j. Thus, $A(z; x_1, \ldots, x_k) - (x_1 + \cdots + x_k)$ counts the trees in $\mathcal{R}$ with more than one vertex (and thus the root being unlabeled). For brevity, we write

$$
\begin{aligned}
H_1 &= \frac{A(z; x_1, \ldots, x_k) - (x_1 + \cdots + x_k)}{z}, \\
H_2 &= A(z; x_1, \ldots, x_k) + H_1 \\
&= \frac{(1+z)A(z; x_1, \ldots, x_k) - (x_1 + \cdots + x_k)}{z}, \text{ and} \\
H_3 &= H_2 + 1 = \frac{(1+z)A(z; x_1, \ldots, x_k) - (x_1 + \cdots + x_k - z)}{z}.
\end{aligned}
$$

In particular, H_1 counts the rooted finite forests that are not just a single tree (i.e. disjoint unions of at least two elements in $\mathcal{R}$), since it counts the objects obtained by removing the unlabeled root of a leaf-labeled tree and rooting each tree of the resulting forest at the neighbor of the old root. Since the neighbors of the old root are either leaves or vertices of degree at least 3, the roots of this forest are either labeled vertices of a singleton or unlabeled vertices of degree at least 2. Thus, all of the trees in the resulting rooted forest are contained in $\mathcal{R}$.

If we take a tree in $\mathcal{R}$ that is not a vertex, its root has degree at least 2. Thus the trees in $\mathcal{R}$ having at least two vertices are in one-to-one correspondence with the rooted forests that have at least two components. Thus H_1 counts the rooted finite forests that have at least two components, and $A(x_1, \ldots, x_k)$ counts the rooted finite forests with precisely one component. Thus, H_2 counts all rooted finite nonempty forests, and H_3 counts all rooted finite forests of trees, including the empty forest.

Any rooted forest (including the empty one) is determined by the number of copies of any tree in $\mathcal{R}$ that appears within it. Therefore H_2 is an infinite sum where each term is of the following form: Let D be a (possibly empty) finite subset of $\mathcal{R}$, for each $T \in D$ let m_T be a positive integer. Then the product $\prod_{T \in D} \mathrm{term}(T)^{m_T}$ is the term corresponding to the forest where each $T \in D$ appears precisely m_T times. Moreover, H_3 is the sum of all terms of this type. Therefore

$$
\begin{aligned}
H_3 &= \left(\prod_{T \in \mathcal{R}} \left(\sum_{j=0}^{\infty} \mathrm{term}(T)^j \right) \right) = \left(\prod_{T \in \mathcal{R}} \left(1 - \mathrm{term}(T) \right)^{-1} \right) \\
&= \left(\prod_{(u; n_1, \ldots, n_k)} (1 - z^u x_1^{n_1} \cdots x_k^{n_k})^{-a_{u; n_1, \ldots, n_k}} \right).
\end{aligned}
$$

The second line follows from collecting the terms corresponding to the trees that have the same form for $\mathrm{term}(T)$ and the definition of the numbers $a_{u; n_1, \ldots, n_k}$. This implies that

$$
\begin{aligned}
\log(H_3) &= - \sum_{(u; n_1, \ldots, n_k)} a_{n_1, \ldots, n_k} \log(1 - z^u x_1^{n_1} \cdots x_k^{n_k}) \\
&= \sum_{(u; n_1, \ldots, n_k)} a_{u; n_1, \ldots, n_k} \sum_{n=1}^{\infty} \frac{(z^u x_1^{n_1} \cdot x_k^{n_k})^n}{n}
\end{aligned}
$$

12

$$= \sum_{n=1}^{\infty} \frac{1}{n} \sum_{(u;n_1,\ldots,n_k)} a_{n_1,\ldots,n_k} ((z^n)^u (x_1^n)^{n_1} \cdot (x_k^n)^{n_k})$$

$$= \sum_{n=1}^{\infty} \frac{1}{n} A(z^n; x_1^n, \ldots, x_k^n),$$

from which the statement in the theorem follows. $\qquad\square$

As an immediate corollary, we can now give a formula involving the generating function for the number of trees in $\mathcal{R}$ where the label j is used precisely n_j times: Let $g_{n_1,\ldots,n_k}$ be the number of such trees in $\mathcal{R}$, put

$$G(x_1,\ldots,x_k) = \sum_{(n_1,\ldots,n_k)} g_{n_1,\ldots,n_k} \prod_{j=1}^{k} x_j^{n_j},$$

and let $z = 1$ in the statement of Theorem 6. Then

Corollary 1.

$$G(x_1,\ldots,x_k) = \frac{1}{2}\left((x_1 + \cdots + x_k - 1) + \mathrm{Exp}\Big(\sum_{n=1}^{\infty} \frac{1}{n} G(x_1^n,\ldots,x_k^n) \Big) \right).$$

We illustrate the use of this formula by deriving a recursion for the number $g_{n;k}$ of trees in $\mathcal{R}$ on n leaves using $[k]$ as label set. Clearly $G_k(x) = \sum_n g_{n;k} x^n = G(x,\ldots,x)$. Put $G_k^*(x) = \sum_{n\geq 1} \frac{1}{n} G_k(x^n) = \sum_{n\geq 0} g_{n;k}^* x^n$. Then $g_{0;k}^* = g_{0;k} = 0$, and for $n \geq 1$ we have

$$g_{n;k}^* = \frac{1}{n} \sum_{d:d|n} d g_{d;k} = g_{n;k} + \frac{1}{n} \sum_{\substack{d:d|n \\ d<n}} d g_{d;k}. \tag{7}$$

Therefore $g_{1;k}^* = g_{1;k}$. From Corollary 1 it follows that

$$G_k(x) = \frac{1}{2}\left(kx + \sum_{m\geq 1} \frac{(G_k^*(x))^m}{m!} \right).$$

In particular, we get $g_{1;k} = \frac{1}{2}(k + g_{1;k})$ (i.e. $g_{1;k} = k$, as expected since $g_{1;k}$ counts the labeled single vertex trees!). Moreover, for $n \geq 2$ we get

$$\begin{aligned} 2g_{n;k} &= \sum_{m=1}^{n} \left(\frac{1}{m!} \sum_{\substack{(n_1,\ldots,n_m):n_i\geq 1 \\ n_1+\cdots+n_m=n}} \prod_{j=1}^{m} g_{n_j;k}^* \right) \\ &= g_{n;k}^* + \sum_{m=2}^{n} \left(\frac{1}{m!} \sum_{\substack{(n_1,\ldots,n_m):n_i\geq 1 \\ n_1+\cdots+n_m=n}} \prod_{j=1}^{m} g_{n_j;k}^* \right), \end{aligned}$$

from which, using (7), we can obtain (for $n \geq 2$) that

$$g_{n;k} = \frac{1}{n} \sum_{\substack{d:d|n \\ d<n}} d g_{d;k} + \sum_{m=2}^{n} \left(\frac{1}{m!} \sum_{\substack{(n_1,\ldots,n_m):n_i\geq 1 \\ n_1+\cdots+n_m=n}} \prod_{j=1}^{m} \left(\frac{1}{n_j} \sum_{d:d|n_j} d g_{d;k} \right) \right). \tag{8}$$

13

Note that rooted non-binary tree shapes (i.e. unlabeled rooted trees where internal non-root vertices have degree at least 3 and the root does not have degree 1) are in one-to-one correspondence with the rooted trees in $\mathcal{R}$ where only the label 1 is used. t Thus, these shapes are counted by (8) using the substitution $k = 1$.

Using Theorem 2, we now obtain analogous results for counting unrooted trees. Let $\mathcal{B}$ denote the class of unrooted leaf-multi-labeled trees where every internal vertex has degree at least 3. Let $b_{u;n_1,\ldots,n_k}$ denote the number of trees in $\mathcal{B}$ that have u unlabeled vertices and in which precisely n_j copies of the label j are used, and put $B(z; x_1, \ldots, x_k) = \sum b_{u;n_1,\ldots,n_k} z^u x_1^{n_1} \cdots x_k^{n_k}$.

To give a formula for the function B in terms of A, it is helpful to slightly extend the definition of p_T given in Section 3. We denote by $p_{T;un}$ the number of nonequivalent, unlabeled points of a leaf-multi-labeled unrooted tree, and by $p_{T;j}$ the number of nonequivalent points of T that are labeled with j. Clearly, $p_T = p_{T;un} + \sum_{j=1}^k p_{T;j}$, and

$$p_T - q_T + s_T \quad = \quad p_{T;un} + \sum_{j=1}^k p_{T;j} - q_T + s_T = 1. \tag{9}$$

Using this we obtain

Theorem 7.

$$\begin{aligned}
B(z; x_1, \ldots, x_k) \quad = \quad & (1 + x_1 + \cdots + x_k) A(z; x_1, \ldots, x_k) \\
& - \frac{1}{2}\Big((z+1) A^2(z; x_1, \ldots, x_k) + (z-1) A(z^2; x_1^2, \ldots, x_k^2) \Big).
\end{aligned}$$

Proof. For brevity, $A(\cdot^m)$ will be used for $A(z^m; x_1^m, \ldots, x_k^m)$. By (9),

$$B(z; x_1, \ldots, x_k) = \sum_{T \in \mathcal{B}} \mathrm{term}(T) = \sum_{T \in \mathcal{B}} \mathrm{term}(T)(p_{T;un} + \sum_{j=1}^k p_{T;j} - q_T + s_T).$$

For any unrooted leaf-multi-labeled tree T, $p_{T;un}$ is the number of trees in $\mathcal{R}$ that are isomorphic to T and whose root is an unlabeled vertex of T (in particular, the root has degree at least 3). In addition, $p_{T;j}$ is the number of leaf-multi-labeled trees that are isomorphic to T and have a leaf-vertex with label j marked; q_T is the number of trees in $\mathcal{R}$ where the root has degree 2 and, after suppressing the root vertex, we obtain a tree that is isomorphic to T; and s_T is the number of trees that are counted by q_T for which the two subtrees at the root are isomorphic.

Now, to obtain the terms of $B(\cdot)$ corresponding to $\sum_T \mathrm{term}(T) \sum_j p_{T;j}$, first note that the contribution of the single vertex trees marked at a (leaf-)vertex is counted by $\sum_j x_j$. Also, the contribution of the trees with at least 2 vertices that are marked at a leaf-vertex is counted by $A(\cdot) \sum_j x_j$, since removing the marked vertex and rooting the remaining tree at the neighbor of this marked vertex gives a tree in $\mathcal{R}$. Thus $\sum_T \mathrm{term}(T) \sum_j p_{T;j} = (A(\cdot) + 1) \sum x_j$.

We now consider the terms corresponding to $\sum_T \mathrm{term}(T) p_{T;un}$. If we consider the unlabeled marked vertex root, we get a tree in $\mathcal{R}$ whose root must have degree at least 3. Also, using similar arguments to those used in the proof of Theorem 1, it can be

14

checked that the trees in $\mathcal{R}$ with root having degree less than 3 (so 2 or 0) are counted by $\frac{z}{2}(A^2(\cdot) + A(\cdot^2)) + \sum_j x_j$, therefore $\sum_T \mathrm{term}(T) p_{T;un} = A(\cdot) - \frac{z}{2}(A^2(\cdot) + A(\cdot^2)) - \sum_j x_j$.

So $\sum_{T \in \mathcal{B}} \mathrm{term}(T)(p_{T;un} + \sum_j p_{T;j}) = (1 + \sum_j x_j)A(\cdot) - \frac{z}{2}(A^2(\cdot) + A(\cdot^2))$.

To complete the proof, note that $\sum_{T \in \mathcal{B}} \mathrm{term}(T)(q_T - s_T)$ counts those rooted leaf-multi-labeled trees (without counting their roots) where the root has degree 2 and the two rooted subtrees obtained when removing the original root are non-isomorphic. Again, using arguments similar to the ones used in Theorem 1 we obtain $\sum_{T \in \mathcal{B}} \mathrm{term}(T)(q_T - s_T) = \frac{1}{2}(A^2(\cdot) - A(\cdot^2))$, as required. $\qquad \square$

We now use this result to give a formula for the generating function for the unrooted leaf-multi-labeled trees without having to keep track of the number of unlabeled vertices: Let $s_{n_1,\dots,n_k}$ denote the unrooted leaf-multi-labeled trees where no vertex has degree 2, and exactly n_j copies of the label j used, and put $S(x_1,\dots,x_k) = \sum s_{n_1,\dots,n_k} x_1^{n_1} \cdots x_k^{n_k}$. Then setting $z = 1$ in the statement of Theorem 7 we obtain

Corollary 2.

$$S(x_1,\dots,x_k) \quad = \quad G(x_1,\dots,x_k)(x_1 + \cdots + x_k + 1) - G^2(x_1,\dots,x_k).$$

Using this in a similar way to that described above for $g_{n;k}$, we obtain a recursion for counting the number $s_{n;k}$ of unrooted leaf-multi-labeled trees on n leaves using $[k]$ as label set:

$$s_{n;k} = \begin{cases} 0 & \text{if } n = 0, \\ k & \text{if } n = 1, \\ kg_{n-1;k} + g_{n;k} - \displaystyle\sum_{j=1}^{n-1} g_{j;k} g_{n-j;k} & \text{if } n \geq 2. \end{cases}$$

As before, $s_{n;1}$ (substitute $k = 1$ in the above formula) counts the unrooted non-binary tree shapes on n leaves (trees where internal vertices have degree at least 3).

We remark that similar formulae can be derived for generating functions and recursions that count the number of leaf-multi-labeled trees in which a specified number of unlabeled, degree 2 vertices are permitted (see [16]).

6. Acknowledgements

7. Role of the funding source

[1] Brualdi R (1999) Introductory Combinatorics. 3rd Edition, Prentice Hall
[2] Cayley A (1889) A theorem on trees, Quart. J. Math. 23:376-378.
[3] Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. American Journal of Human Genetics 19:233-257
[4] Dobson AJ (1974) Unrooted trees for numerical taxonomy. Journal of Applied Probability 11:32-42
[5] Fellows M, Hallett M, Stege U (2003) Analogs & duals of the MAST problem for sequences & trees. Journal of Algorithms 49:192-216
[6] Felsenstein J (1978) The number of evolutionary trees. Systematic Zoology 27:27-33
[7] Flajolet P (1997) A problem in statistical classification theory. http://algo.inria.fr/libraries/autocomb/schroeder-html/schroeder.html
[8] Ganapathy G, Goodson B, Jansen R, Le HS, Ramachandran V, Warnow T (2006) Pattern identification in biogeography. IEEE Transactions on Computational Biology and Bioinformatics 3:334-346
[9] Guillemot S, Jansson J, Sung W (2011) Computing a smallest leaf-multi-labeled phylogenetic tree from rooted triplets. IEEE/ACM Trans Comput Biol Bioinform.(4):1141-7.
[10] Harding EF (1971) The probabilities of rooted tree shapes generated by random bifurcation. Advances in Applied Probability 3:44-77
[11] Harary F, Palmer EM (1973) Graphical Enumeration Academic Press, Inc. (London) LTD.
[12] Harary F, Prins G (1959) The number of homeomorphically irreducible trees, and other species. Acta Mathematica, 101:141-162
[13] Huber K, Oxelman B, Lott M, Moulton V (2006) Reconstructing the evolutionary history of polyploids from multilabeled trees, *Molecular Biology and Evolution*, 23:1784-1791
[14] Huber K, Lott M, Moulton V, Spillner A (2008) The complexity of deriving multilabeled trees from bipartitions. Journal of Computational Biology, 15(6):639-651
[15] Huber K, Moulton V (2006) Phylogenetic networks from multi-labelled trees. Journal of Mathematical Biology 52(5):613-632
[16] V. Johnson (2012) Enumeration Results on leaf-labeled trees. Ph.D. thesis, Department of Mathematics, University of South Carolina, Columbia
[17] Kirchoff G (1847) über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wird. Ann. Phys. Chem. 72:497-508
[18] Lott M, Spillner A, Huber KT, Petri A, Oxelman B, Moulton V (2009) Inferring polyploid phylogenies from multiply-labeled gene trees. BMC Evolutionary Biology 9:216
[19] Lovász L (1993) Combinatorial Problems and Exercises. Chapter 4. North–Holland
[20] Moon JW (1970) Counting Labelled Trees. Canad. Math. Monographs No. 1
[21] Otter R (1948) The number of trees. Annals of Mathematics 49:583-599
[22] Schröder E (1870) Vier combinatorische probleme. Zeitschrift für Mathematik und Physik 15:361-376
[23] Scornavacca C, Berry V, Ranwez V (2009) Building species trees from larger parts of phylogenomic databases. Information and Computation, (209) : pp. 590-605.
[24] Semple C, Steel M (2003) Phylogenetics. Oxford University Press
[25] Wedderburn JHM (1923) The functional equation "$g(x^2) = 2ax + [g(x)]^2$". Annals of Mathematics, 24:121-140