ENumeration of rooted trees and forests

Lajos Takács,* Case Western Reserve University

Abstract

The aim of this paper is to show that we can enumerate various rooted trees and forests by making use of a simple combinatorial theorem which is a generalization of the classical ballot theorem of J. Bertrand.

Counting trees and forests

AMS 1991 Subject Classification: Primary 05C30

1. Introduction

The problem of enumerating various sets of trees and forests arises in many fields such as computer science (sorting and searching data), neuroscience (neural network of the brain), and the molecular study of genes (genesis of humans). In the last three decades, after a long delay which followed the pioneering work of Cayley in 1857, several types of trees and forests have been enumerated, mostly by lengthy analytical calculations. The aim of this paper is to give a simple, elementary and general method which can be applied to any sets of labeled or unlabeled trees or forests of interest. A general theorem is proved and used to derive some known and some new formulas for trees and forests. Our approach is based on a new representation of trees and forests stemming from the theory of queues. The advantage of generating a set of trees or forests by suitably chosen queuing processes is that for the queuing processes we can apply a simple combinatorial theorem which is a generalization of the classical ballot theorem of Bertrand [1].

2. Unlabeled trees

Define $S_n$ as the set of all the rooted, oriented (plane) trees with $n$ unlabeled vertices. There are $|S_4| = 5$ different trees with four vertices, as shown in Figure 1. The root of a tree is a vertex distinguished from the other vertices (depicted in black). The branches are ordered. By interchanging the locations of two different branches, relative to the root, we obtain two different trees. As we shall see, if $n \geq 1$, the number of trees in $S_n$ is

$$|S_n| = \binom{2n-2}{n-1} \frac{1}{n}.$$  \hfill (1)

Received 2 October 1992; revision received 30 December 1992.
*Postal address: 2410 Newbury Drive, Cleveland Heights, OH 44118, USA.
Figure 1. Rooted oriented trees with four unlabeled vertices

In the set $S_n$ there are no restrictions on the degrees of the vertices of the trees. However, in many applications we encounter rooted trees in which the degrees of the vertices are subject to certain constraints. For example, in computer science, trivalent rooted trees play an important role. In a trivalent rooted tree every vertex has degree 3, except the root which has degree 2 and the end vertices which have degree 1.

To study various sets of rooted trees in the case where the degrees of the vertices satisfy some constraints, let us define $R$ as a fixed set of non-negative integers which always contains 0. Denote by $S_n(R)$ the subset of $S_n$ which contains all the trees in $S_n$ in which the degree of the root is in $R$ and, if $j$ is the degree of any other vertex of a tree, then $j-1 \in R$.

**Theorem 1.** The number of trees in $S_n(R)$ is

$$|S_n(R)| = \frac{1}{n} \text{Coeff. of } x^{n-1} \text{ in } \left( \sum_{i \in R} x^i \right)^n. \quad (2)$$

We shall prove later that (2) is an immediate consequence of Theorem 4.

If $R = \{0, 1, 2, \ldots\}$, then $S_n(R) = S_n$ and by (2)

$$|S_n| = \frac{1}{n} \text{Coeff. of } x^{n-1} \text{ in } (1-x)^{-n}. \quad (3)$$

Hence (1) follows.

If $R = \{0, 2\}$, then necessarily $n = 2m + 1$ ($m = 0, 1, 2, \ldots$) and by (2)

$$|S_n(R)| = \binom{2m}{m} \frac{1}{m+1}. \quad (4)$$

Formula (4) gives the number of distinct, trivalent, rooted trees with $2m+1$ unlabeled vertices. Formula (4) was proved in 1859 by Cayley [3]. Harary et al. [10] demonstrated that the number of distinct, oriented, rooted trees with $m+1$ unlabeled vertices is also given by Cayley’s formula (4), that is, if in (1) $n$ is replaced by $m+1$ we get (4). See also De Bruijn and Morselt [6] and Klarner [14], [15].

If $R = \{0, r\}$, where $r \geq 1$, then necessarily $n = rm + 1$ ($m = 0, 1, 2, \ldots$) and by (2)

$$|S_n(R)| = \binom{rm}{m} \frac{1}{m(r-1)+1}. \quad (5)$$
If $R = \{0, 1, 2, \ldots, r\}$, where $r \geq 1$, then by (2)
\[ |S_n(R)| = \frac{1}{n} \sum_{j=0}^{(n-1)/(r+1)} (-1)^j \binom{n}{j} \binom{2n-2-j(r+1)}{n-1}. \]  
(6)

We can also express formula (6) as
\[ |S_n(R)| = \frac{(r+1)^n}{n} P_{2n-1}(n, r+1), \]  
(7)

where $P_l(n, r+1)$ can be interpreted in the following way. A box contains $r+1$ cards numbered 1, 2, ..., $r+1$. We draw $n$ cards with replacement. Then $P_l(n, r+1)$ is the probability that the sum of the numbers drawn is $l$ provided that all the possible results are equally probable. The probability $P_l(n, r+1)$ for $n \leq l \leq n(r+1)$ was found in 1710 by Montmort [17] and in 1730 by De Moivre [7]. For the history of this problem see Takács [28].

If $R = \{0, 2, 4, 6, \ldots\}$, then necessarily $n = 2m+1$ ($m = 0, 1, 2, \ldots$) and by (2)
\[ |S_n(R)| = \binom{3m}{m} \frac{1}{2m+1}. \]  
(8)

3. Labeled trees

Denote by $S_n^*$ the set of all rooted trees with $n$ labeled vertices. See Figure 2 for $S_n^*$. There are $|S_n^*| = 64$ rooted trees with four labeled vertices, 1, 2, 3, 4. In Figure 2 the roots of the trees are depicted in black and only one tree is displayed in each of four subsets of $S_n^*$: (a), (b), (c) and (d). The sets (a), (b), (c) and (d) contain 24, 12, 24 and 4 trees, respectively. As we shall see, the number of trees in $S_n^*$ is
\[ |S_n^*| = n^{n-1} \]  
(9)

if $n \geq 1$.

In the set $S_n^*$ there are no restrictions on the degrees of the vertices of the trees. To study various subsets of $S_n^*$ in the case where the degrees of the vertices satisfy some constraints, let us define $R$ as a fixed set of non-negative integers which always contains 0. Denote by $S_n^*(R)$ the subset of $S_n^*$ which contains all the trees in $S_n^*$ in which the degree of the root is in $R$ and, if $j$ is the degree of any other vertex of a tree, then $j-1 \in R$. 

Figure 2. Rooted trees with four labeled vertices
**Theorem 2.** The number of trees in \( S^*_n(R) \) is

\[
|S^*_n(R)| = (n-1)! \text{ Coeff. of } x^{n-1} \text{ in } \left( \sum_{i \in R} \frac{x^i}{i!} \right)^n. \tag{10}
\]

We shall demonstrate later that (10) is an immediate consequence of Theorem 4.

If \( R = \{0, 1, 2, \ldots\} \), then \( S^*_n(R) = S^*_n \) and by (10)

\[
|S^*_n| = (n-1)! \text{ Coeff. of } x^{n-1} \text{ in } e^{ax}. \tag{11}
\]

Hence (9) follows. Formula (9) was discovered by Cayley [4] in 1889. Since then various proofs have been given for (9) by Dziobek [9], Prüfer [22], Bol [2], Clarke [5], Moon [18], [19], Rényi [23], [24], Takács [30], [31], and others. It should be added that Cayley's formula (9) can also be derived from an 1847 result of Kirchhoff [11], [12] for spanning trees. It seems the above proof is the simplest possible proof for Cayley's formula (9).

If \( R = \{0, 2\} \), then necessarily \( n = 2m+1 \) and by (10)

\[
|S^*_n(R)| = \binom{2m+1}{m} \binom{2m}{m}^{2^m}. \tag{12}
\]

If \( R = \{0, r\} \) and \( r \geq 1 \), then necessarily \( n = rm+1 \ (m = 0, 1, 2, \ldots) \) and by (10)

\[
|S^*_n(R)| = \binom{rm+1}{m} \binom{rm}{m}^{m m}. \tag{13}
\]

If \( R = \{0, 1, \ldots, r\} \) and \( r \geq 1 \), then by (10)

\[
|S^*_n(R)| = (n-1)! \text{ Coeff. of } x^{n-1} \text{ in } \left( \sum_{i=0}^{r} \frac{x^i}{i!} \right)^n. \tag{14}
\]

In this case we can write that

\[
|S^*_n(R)| = n^{n-1} P(\xi_n(r+1) \geq n), \tag{15}
\]

where the random variable \( \xi_n(r+1) \) is connected with the generalized birthday problem. To define \( \xi_n(r+1) \) let us put balls at random one by one in \( n \) boxes until one of the boxes contains \( r+1 \) balls. It is assumed that every box has the same probability. The random variable \( \xi_n(r+1) \) is defined as the number of balls needed. The distribution of \( \xi_n(r+1) \) has been studied by Klamkin and Newman [13], and Dwass [8].

If \( R = \{0, 2, 4, 6, \ldots\} \), then necessarily \( n = 2m+1 \ (m = 0, 1, 2, \ldots) \) and by (10)

\[
|S^*_n(R)| = \frac{1}{2^n} \sum_{j=0}^{n} \binom{n}{j}(2j-n)^{n-1}. \tag{16}
\]

We can express (16) also as
where $v_n$ is a random variable which has a Bernoulli distribution with parameters $n$ and $p = \frac{1}{2}$. If $n = 2m + 1$, we have

$$|S_n^*(R)| = \sqrt{2\pi(2m)^{2m}e^{-m}},$$

as $m \to \infty$.

4. Representation of trees

The proofs of Theorems 1 and 2 are based on a new representation of trees.

We defined $S_n$ as the set of all the rooted, oriented (plane) trees with $n$ unlabeled vertices. Equivalently, we can define $S_n$ as the set of sequences $(i_1, i_2, ..., i_n)$ where $i_1, i_2, ..., i_n$ are non-negative integers which satisfy the conditions

$$i_1 + i_2 + \cdots + i_n = n - 1$$

and

$$i_1 + i_2 + \cdots + i_r \geq r \quad (1 \leq r \leq n - 1).$$

If $R$ is a fixed set of non-negative integers, then let us denote by $S_n(R)$ the set of sequences $(i_1, i_2, ..., i_n) \in S_n$ in which $i_r \in R$ for all $r = 1, 2, ..., n$.

With every sequence $(i_1, i_2, ..., i_n)$ in $S_n$ we associate a rooted tree. The tree has vertex set $(1, 2, ..., n)$ and vertex 1 is designated as the root of the tree. Two vertices $r$ and $s$ $(1 \leq r < s \leq n)$ are joined by an edge if and only if

$$i_0 + i_1 + \cdots + i_{r-1} < s \leq i_0 + i_1 + \cdots + i_r,$$

where $i_0 = 1$. In the tree $(i_1, i_2, ..., i_n)$, the root has degree $i_1$ and vertex $r$ $(1 < r \leq n)$ has degree $i_r + 1$.

If $(i_1, i_2, ..., i_n) \in S_n(R)$, then the degrees of the vertices are subject to the constraints $i_r \in R$ for $r = 1, 2, ..., n$.

Actually, in the above definition of $S_n(R)$ we assign labels to the vertices of the trees, but if we are not interested, the labels can be disregarded and each tree can be visualized as an unlabeled tree. We can interpret $|S_n(R)|$ as the number of unlabeled distinct rooted trees in the case where the degrees of the vertices satisfy the constraints imposed by $R$. However, for many purposes it is convenient to retain the labels assigned to the vertices. For example, if we want to generate and study trees by a computer program then it is natural to generate the set $S_n(R)$ and if we are not interested we can ignore the labels.

The above representation of trees stems from the theory of queues. We represent a tree with vertex set $(1, 2, ..., n)$ by a single server queue. It is supposed that initially, when the server starts working, the first customer is already waiting for service. The server serves this customer and all the new arrivals as long as they keep coming. Denote by $i_1, i_2, ...$ the number of arrivals during the first, second, ... service times, respectively. If there are no more customers to serve, the initial busy period ends. If the initial busy period consists of $n$ services, let us associate the following graph with the queuing
process considered: the graph has vertex set \((1, 2, \ldots, n)\) and vertices \(r\) and \(s\), where \(1 \leq r < s \leq n\), are joined by an edge if and only if the \(s\)th customer arrives during the service time of the \(r\)th customer. Evidently, the graph is a tree with vertex set \((1, 2, \ldots, n)\). Different queuing processes yield different trees.

Let us choose a tree \((i_1, i_2, \ldots, i_n)\) in \(S_n(R)\). We can label the vertices of \((i_1, i_2, \ldots, i_n)\) in

\[
\frac{n!}{i_1! i_2! \cdots i_n!}
\]

(22)
different ways. If we perform all the possible labelings on all the trees in \(S_n(R)\), we get a set of labeled, distinct trees, which we shall denote by \(S_n^*(R)\). If \(R = \{0, 1, 2, \ldots\}\), we write \(S_n^*(R) = S_n^*\). These definitions of \(S_n^*\) and \(S_n^*(R)\) are in agreement with the earlier ones.

Accordingly, \(|S_n(R)|\) is the number of sequences \((i_1, i_2, \ldots, i_n)\) of non-negative integers which satisfy the conditions (19), (20) and \(i_r \in R\) for \(r = 1, 2, \ldots, n\). Furthermore,

\[
|S_n^*(R)| = \sum_{(i_1, i_2, \ldots, i_n) \in S_n(R)} \frac{n!}{i_1! i_2! \cdots i_n!}.
\]

(23)

5. Combinatorial theorems

We can determine \(|S_n(R)|\) and \(|S_n^*(R)|\) in a simple way by using combinatorial methods.

**Theorem 3.** Let us suppose that \(n\) cards are marked with non-negative integers \(k_1, k_2, \ldots, k_n\), respectively, where \(k_1 + k_2 + \cdots + k_n = k \leq n\). Among the \(n!\) permutations of the \(n\) cards there are exactly

\[
S(n, k) = (n-k)(n-1)!
\]

(24)
permutations in which the sum of the numbers on the first \(r\) cards is less than \(r\) for every \(r = 1, 2, \ldots, n\).

**Proof.** We can prove by mathematical induction that \(S(n, k)\) does not depend individually on \(k_1, k_2, \ldots, k_n\), it depends only on their sum \(k\) and their number \(n\), and is given by (24). Obviously, \(S(1, 0) = 1\) and \(S(1, 1) = 0\). Let us suppose that \(S(m, k) = (m-k)(m-1)!\) for \(0 \leq k \leq m \leq n-1\), where \(n \geq 2\). If we take into consideration that the last card in the \(n!\) permutations of the \(n\) cards may be marked \(k_1, k_2, \ldots, k_n\), then we can write down that

\[
S(n, k) = \sum_{i=1}^{n} S(n-1, k-k_i)
\]

(25)
for \(k < n\) and \(S(n, n) = 0\). If \(k < n\), then by the induction hypothesis

\[
S(n, k) = \sum_{i=1}^{n} (n-1-k+k_i)(n-2)! = (n-k)(n-1)!.\]

(26)
Consequently, (24) is true for all \(n = 1, 2, \ldots\) and \(0 \leq k \leq n\).
We can prove also that among the $n$ cyclic permutations of the $n$ cards there are $n - k$ cyclic permutations in which the sum of the numbers on the first $r$ cards is less than $r$ for every $r = 1, 2, \ldots, n$. This too implies (24). Theorem 3 is a generalization of the classical ballot theorem of Bertrand [1]. See Takács [27], [29].

The number of ways in which a non-negative integer $k$ can be represented as a sum of $n$ $(n \geq 1)$ non-negative integers is

$$\binom{n+k-1}{n-1},$$

that is, (27) is the number of solutions of the equation

$$k_1 + k_2 + \cdots + k_n = k$$

in non-negative integers. Obviously, $(k_1, k_2, \ldots, k_n)$ is a solution if and only if $(k_1 + 1, k_2 + 2, \ldots, k_{n-1} + n - 1)$ is a combination without repetition of size $n - 1$ of the elements $1, 2, \ldots, n + k - 1$.

If we require that the solution $(k_1, k_2, \ldots, k_n)$ of (28) satisfy also

$$k_1 + k_2 + \cdots + k_r < r \quad (1 \leq r \leq n),$$

then by Theorem 3 we obtain that there are

$$\binom{n+k-1}{n-1} \frac{n-k}{n}$$

such sequences $(k_1, k_2, \ldots, k_n)$.

More generally, we have the following identity.

**Theorem 4.** Let $f(k_1, k_2, \ldots, k_n)$ be a symmetric function of the variables $k_1, k_2, \ldots, k_n$. Then

$$\sum_{k_1 + k_2 + \cdots + k_n = k \atop k_1 + \cdots + k_r < r \quad (1 \leq r \leq n)} f(k_1, k_2, \ldots, k_n) = \frac{n-k}{n} \sum_{k_1 + k_2 + \cdots + k_n = k} f(k_1, k_2, \ldots, k_n)$$

for $0 \leq k \leq n$.

**Proof.** We can prove (31) by the repeated applications of Theorem 3 or by mathematical induction on $n$. We note that (31) still remains valid if we assume only that the function $f(k_1, k_2, \ldots, k_n)$ is invariant under the $n$ cyclic permutations of $(k_1, k_2, \ldots, k_n)$. If, in particular,

$$f(k_1, k_2, \ldots, k_n) = g(k_1)g(k_2)\cdots g(k_n),$$

then Theorem 4 is applicable and in (31)

$$\sum_{k_1 + k_2 + \cdots + k_n = k} f(k_1, k_2, \ldots, k_n) = \text{Coeff. of } x^k \text{ in } \left( \sum_{i=0}^{\infty} g(i)x^i \right)^n.$$

If in (31), (32) and (33), $k = n - 1$, $k_r = i_{n+1-r}$ for $1 \leq r \leq n$, $g(i) = 1$ for
$i \in R$ and $g(i) = 0$ for $i \notin R$, then we obtain (2), and if $k = n-1$, $k_r = i_{n+1-r}$ for $1 \leq r \leq n$, $g(i) = 1/i!$ for $i \in R$ and $g(i) = 0$ for $i \notin R$, then we obtain (10).

6. Forests

A forest is a simple graph that has no cycles. In other words, a forest is a simple graph, all of whose components are trees. In the same way as we defined rooted trees we can define forests consisting of $p$ rooted trees. Denote by $F_{n,p}$ ($1 \leq p \leq n$) the set of sequences $(i_1,i_2,\ldots,i_n)$, where $i_1,i_2,\ldots,i_n$ are non-negative integers which satisfy the conditions

$$i_1+i_2+\cdots+i_n = n-p \quad (34)$$

and

$$i_1+i_2+\cdots+i_r > r-p \quad (p \leq r \leq n-1). \quad (35)$$

If $R$ is a fixed set of non-negative integers, then we denote by $F_{n,p}(R)$ the set of sequences $(i_1,i_2,\ldots,i_n) \in F_{n,p}$ for which $i_r \in R$ for all $r = 1,2,\ldots,n$.

With every sequence $(i_1,i_2,\ldots,i_n)$ in $F_{n,p}$ we associate a forest which has vertex set $(1,2,\ldots,n)$, and consists of $p$ rooted trees whose roots are labeled $1,2,\ldots,p$. Two vertices, $r$ and $s$ ($1 \leq r < s \leq n$) are joined by an edge if and only if

$$i_0+i_1+\cdots+i_{r-1} < s \leq i_0+i_1+\cdots+i_r, \quad (36)$$

where $i_0 = p$. In the forest $(i_1,\ldots,i_n) \in F_{n,p}$ the roots of the trees, $1,2,\ldots,p$, have degrees $i_1,i_2,\ldots,i_p$, respectively, and vertex $r$ ($p < r \leq n$) has degree $i_r+1$. If $(i_1,i_2,\ldots,i_n) \in F_{n,p}(R)$, then the degrees of the vertices are subject to the constraints $i_r \in R$ for $r = 1,2,\ldots,n$.

The number of forests in $F_{n,p}$ is

$$|F_{n,p}| = \frac{p}{n} \binom{2n-p-1}{n-1}. \quad (37)$$

This follows from (31) and (32), where now $k = n-p$, $k_r = i_{n+1-r}$ for $r = 1,2,\ldots,n$ and $g(i) = 1$ for $i = 0,1,2,\ldots$. If we consider a forest $(i_1,i_2,\ldots,i_n)$ in $F_{n,p}(R)$, we can label the vertices of $(i_1,i_2,\ldots,i_n)$ in

$$\frac{n!}{i_1!i_2!\cdots i_n!} \quad (38)$$

ways. If we perform all the possible labelings on all the forests in $F_{n,p}(R)$, we get a set of labeled forests $F_{n,p}^*(R)$. If $R = \{0,1,2,\ldots\}$, we write $F_{n,p}^*(R) = F_{n,p}^*$. The number of forests in $F_{n,p}^*$ is

$$|F_{n,p}^*| = \sum_{F_{n,p}} \frac{n!}{i_1!i_2!\cdots i_n!} = n(n-1)\cdots(n-p+1)pn^{n-p-1} \quad (39)$$

for $1 \leq p \leq n$. This follows from (31) and (32), where now $k = n-p$, $k_r = i_{n+1-r}$ for $r = 1,2,\ldots,n$ and $g(i) = 1/i!$ for $i = 0,1,2,\ldots$. 

By (39) the number of forests having vertex set \( (1, 2, \ldots, n) \) and consisting of \( p \) trees such that vertices \( 1, 2, \ldots, p \) all belong to different trees is

\[
|F_{n, p}^*| = \frac{(n-p)!}{n!} = pn^{n-p-1}
\]  

(40)

for \( 1 \leq p \leq n \).

In 1889 Cayley [4] stated formula (40), but he did not indicate how to prove it. In 1959 Rényi [23] gave an analytic proof for (40). For other proofs of (40) we refer to Moon [20], [21], Göbel [21], p. 17, Riordan [25], Kolchin [16], pp. 150–151, Sachkov [26], pp. 181–183, and Takács [30], [31].

By Theorem 4 we can also determine \(|F_{n, p}(R)|\) and \(|F_{n, p}^*(R)|\) for any set \( R \).

**Theorem 5.** We have

\[
|F_{n, p}(R)| = \frac{p}{n} \text{Coeff. of } x^{n-p} \text{ in } \left( \sum_{i \in R} x_i \right)^n.
\]  

(41)

This follows from (31) and (32), where now \( k = n-p \), \( k_r = i_{n+1-r} \) for \( 1 \leq r \leq n \), \( g(i) = 1 \) if \( i \in R \) and \( g(i) = 0 \) if \( i \notin R \).

**Theorem 6.** We have

\[
|F_{n, p}^*(R)| = p(n-1)! \text{Coeff. of } x^{n-p} \text{ in } \left( \sum_{i \in R} \frac{x_i}{i!} \right)^n.
\]  

(42)

This follows from (31) and (32), where now \( k = n-p \), \( k_r = i_{n+1-r} \) for \( 1 \leq r \leq n \), \( g(i) = \frac{1}{i!} \) if \( i \in R \) and \( g(i) = 0 \) if \( i \notin R \).

The above representation of forests also stems from the queuing model mentioned earlier. If in the queuing process initially \( p \geq 1 \) customers are waiting for service and if the initial busy period consists of \( n \) services, then the corresponding graph is a forest consisting of \( n \) vertices and \( p \) trees.

**References**


