The multinomial distribution on rooted labeled forests*

Jim Pitman

Technical Report No. 499

Department of Statistics
University of California
367 Evans Hall # 3860
Berkeley, CA 94720-3860

June 15, 1998

Abstract

For a probability distribution $(p_s, s \in S)$ on a finite set S, call a random forest $\mathcal{F}$ of rooted trees labeled by S (with edges directed away from the roots) a *p-forest* if given $\mathcal{F}$ has m edges the vector of out-degrees of vertices of $\mathcal{F}$ has a multinomial distribution with parameters m and $(p_s, s \in S)$, and given also these out-degrees the distribution of $\mathcal{F}$ is uniform on all forests with the given out-degrees. The family of distributions of p-forests is studied, and shown to be closed under various operations involving deletion of edges. Some related enumerations of rooted labeled forests are obtained as corollaries.

1 Introduction

Let $\mathbf{F}(S)$ denote the set of all forests of rooted trees labeled by a finite set S of size $|S|$. Each $\mathbf{f} \in \mathbf{F}(S)$ is a directed graph labeled by S, that is a subset of $S \times S$, such that each

*Research supported in part by N.S.F. Grant DMS97-03961

connected component of the graph is a tree with edges directed away from some root vertex. The notation $v \xrightarrow{\mathbf{f}} w$ will be used instead of $(v, w) \in \mathbf{f}$ to show that (v, w) is a directed edge of $\mathbf{f}$. For $s \in S$ and $\mathbf{f} \in \mathbf{F}(S)$ let $\mathbf{f}_s := \{t \in S : s \xrightarrow{\mathbf{f}} t\}$, the *set of children of s in* $\mathbf{f}$. Note that for each forest $\mathbf{f}$ the $\mathbf{f}_s$ are disjoint subsets of S as s ranges over S. The *number of children* or *out-degree* of s in the forest $\mathbf{f}$ is $|\mathbf{f}_s|$. The number of edges of $\mathbf{f}$ is $|\mathbf{f}| = \sum_s |\mathbf{f}_s|$, and the number of tree components of $\mathbf{f}$ is $|S| - |\mathbf{f}|$. The starting point of this paper is the observation of [15] that for each probability distribution $p = (p_s, s \in S)$ on S, and each $1 \le m \le |S| - 1$, the formula

$$P(\mathcal{F} = \mathbf{f}) = \binom{|S|-1}{m}^{-1} \prod_{s \in S} p_s^{|\mathbf{f}_s|} \qquad (\mathbf{f} \in \mathbf{F}(S) : |\mathbf{f}| = m) \tag{1}$$

defines the probability distribution of a random forest $\mathcal{F}$ with m edges. This is a probabilistic expression of the following *multinomial expansion over forests* [15, 18, 21], which is an identity of polynomials in variables $x_s, s \in S$ generalizing Cayley's multinomial expansion over trees [5, 19, 16]:

$$\sum_{\mathbf{f} \in \mathbf{F}(S) : |\mathbf{f}| = m} \prod_{s \in S} x_s^{|\mathbf{f}_s|} = \binom{|S|-1}{m} \left(\sum_{s \in S} x_s \right)^m . \tag{2}$$

Definition 1 For a probability distribution p on S, and $1 \le m \le |S| - 1$, call a random forest $\mathcal{F}$ with distribution (1) a *p-forest with m edges*, or *a p-forest of k trees*, where $k = |S| - m$. Call $\mathcal{F}$ a *p-tree* if $k = 1$. Call a random forest $\mathcal{F}$ a *p-forest* if $\mathcal{F}$ given $|\mathcal{F}| = m$ is a p-forest with m edges for each $1 \le m \le |S| - 1$.

Put another way, a random element $\mathcal{F}$ of $\mathbf{F}(S)$ is a p-forest if and only if the distribution of $\mathcal{F}$ is given by the formula

$$P(\mathcal{F} = \mathbf{f}) = w_{|\mathbf{f}|} \prod_{s \in S} p_s^{|\mathbf{f}_s|} \qquad (\mathbf{f} \in \mathbf{F}(S)) \tag{3}$$

for some sequence of weights $(w_m, 1 \le m \le |S| - 1)$. If p is uniform on S, a p-forest with m edges has uniform distribution on the set of all rooted forests labeled by S with m edges. Many exact combinatorial results and asymptotic distributions are known in this case. See [15] for a review of such results and their applications to random graphs. Here attention is restricted to exact distributional results for p-forests for a general underlying probability distribution p. The main point is to present some properties of p-forests which might prove useful in a variety of contexts. This study was suggested by recent applications of p-forests to the construction of partition-valued and measure-valued coalescent processes [15, 6, 1]. See [16] regarding the connection between p-forests and the model of [4, 10, 20], for a random mapping from S to S with independent images with distribution p, and [16, 17] for the relation between p-forests and random subsets with distributions generated by Hurwitz's [9] binomial expansions. See also [2, 11, 14, 12] concerning other models of random trees and forests and their applications.

The following characterization of of a p-forest follows easily from Definition 1. Here and throughout the paper, the notation $(x)_m := \prod_{i=0}^{m-1}(x-i)$ is used for falling factorials.

Proposition 2 [16] *A random element $\mathcal{F}$ of $\mathbf{F}(S)$ is a p-forest if and only if both*
(i) *for each $1 \le m \le |S|-1$, the conditional distribution of the out-degree count vector $(|\mathcal{F}_s|, s \in S)$ given $|\mathcal{F}| = m$ is multinomial with parameters m and $(p_s, s \in S)$, and*
(ii) *for each vector of counts $(f_s, s \in S)$ with $\sum_s f_s = m$, the conditional distribution of $\mathcal{F}$ given $(|\mathcal{F}_s| = f_s$ for all $s \in S)$ is uniform over the set of $(|S|-1)_m/(\prod_{s\in S} f_s!)$ forests with the given out-degrees.*

For any random rooted forest $\mathcal{F}$ labeled by S with a fixed number m of edges, the vector of out-degree counts $(|\mathcal{F}_s|, s \in S)$ is subject to the constraint $\sum_s |\mathcal{F}_s| = m$. Therefore, the expectation of $|\mathcal{F}_s|$ equals mp_s for some probability distribution p on S. By the previous proposition, for any given p and m this is achieved by a p-forest with m edges. The paper [16] recorded some basic features of p-forests, such as the distribution of the random set of roots of a p-forest of k trees, and the conditional distribution of a p-forest given its set of roots. In particular, the root R of a p-tree $\mathcal{T}$ has distribution p, and R is independent of the unrooted tree derived from $\mathcal{T}$. Several natural constructions of a p-tree for general p are reviewed in [16, §3]. Starting from a p-tree, one construction of a p-forest is given by the following proposition:

Proposition 3 [15] *A p-forest of k trees is obtained by deleting $k-1$ edges picked uniformly at random from the $|S|-1$ edges of a p-tree.*

The main results of this paper are the following three theorems, each of which describes a different way in which the family of distributions of p-forests is closed under operations involving deletion of edges. For a forest $\mathbf{f} \in \mathbf{F}(S)$ and a subset B of S, the *restriction of* $\mathbf{f}$ *to* B is the forest $\mathbf{f}^B \in \mathbf{F}(B)$ defined by $\mathbf{f}^B := \mathbf{f} \cap (B \times B)$. For a probability distribution p on S and a subset B of S, let $p_B := \sum_{s\in B} p_s$. For B with $p_B > 0$, let $p(\cdot \mid B)$ denote the probability distribution on B obtained by conditioning p on B.

Theorem 4 *Let p be a probability distribution on S with $0 \notin S$, let $0 < p_0 < 1$, and let p' be the probability distribution on $\{0\} \cup S$ defined by $p'_0 = p_0$ and $p'_s = (1-p_0)p_s$ for $s \in S$, so $p = p'(\cdot \mid S)$. Let $\mathcal{T}'$ be a p'-tree labeled by $\{0\} \cup S$ and conditioned to have root 0, and let $\mathcal{F}$ be the restriction of $\mathcal{T}'$ to S. Then $\mathcal{F}$ is a p-forest with the same distribution as if each edge of a p-tree were deleted independently with probability p_0.*

This result is easily verified by direct calculation, or by application of [16, Th. 23] with $0 \cup S$ substituted for S and $R = \{0\}$. The next theorem is proved in Section 2:

Theorem 5 (Projection rule for p-forests) *For B a non-empty subset of S and $\mathcal{F}$ a p-forest labeled by S, the restriction $\mathcal{F}^B$ of $\mathcal{F}$ to B is a $p(\cdot \mid B)$-forest. The distribution of $|\mathcal{F}^B|$ on $0, \ldots, |B|-1$ is determined by p_B and the distribution of $|\mathcal{F}|$ via the falling factorial moments*

$$E(|\mathcal{F}^B|)_r = \frac{E(|\mathcal{F}|)_r}{(n-1)_r}(|B|-1)_r\, p_B^r \qquad (r = 0, 1, 2, \ldots). \tag{4}$$

To be explicit, these factorial moments determine the distribution of $|\mathcal{F}^B|$ via the sieve formula [3, p. 17]:

$$P(|\mathcal{F}^B| = \ell) = \sum_{r=\ell}^{|B|-1} \binom{r}{\ell}(-1)^{r-\ell}\frac{E(|\mathcal{F}^B|)_r}{r!} \qquad (0 \le \ell \le |B|-1). \tag{5}$$

Corollary 6 (Projection rule for uniform forests) *Suppose that $\mathcal{F}$ has uniform distribution on the set of all $\binom{|S|-1}{k-1}|S|^{|S|-k}$ forests of k rooted trees labeled by S, for some $1 \le k \le |S|$. Then for each non-empty subset B of S the conditional distribution of $\mathcal{F}^B$ given that $\mathcal{F}^B$ has j components is uniform on the set of all forests of j rooted trees labeled by B. That is to say, each forest $\mathbf{f} \in \mathbf{F}(B)$ with j tree components is the restriction to B of the same number of forests in $\mathbf{F}(S)$ with k tree components.*

This number of forests, which depends only on $|B|$, $|S|$, j and k, can be read from (4) and (5) with $p_B = |B|/|S|$. Underlying the above results is a simple formula, presented in Section 3, for the probability that a p-forest contains a specified set of edges. A straightforward calculation with this formula yields easily the following generalization of Proposition 3:

Theorem 7 *Suppose that $\mathcal{F}$ is p-forest. Given $\mathcal{F}$, let each edge $s \xrightarrow{\mathcal{F}} t$ be marked red with probability r_s, indepedently as (s,t) ranges over all directed edges of $\mathcal{F}$. Let $\mathcal{F}_{\rm red}$ denote the forest of red edges so obtained, and let $p_* := \sum_{s \in S} p_s r_s$. Then $\mathcal{F}_{\rm red}$ is a p'-forest, where $p'_s := p_s r_s / p_*$, and given $\mathcal{F}$ has m edges the number of edges of $\mathcal{F}_{\rm red}$ has a binomial(m, p_*) distribution.*

In particular, if $\mathcal{F}$ is a random tree with uniform distribution on the set of all rooted trees labeled by S, then $\mathcal{F}_{\rm red}$ obtained by the above construction is a p'-forest with p'_s proportional to r_s.

2 The Projection Rule.

This section establishes a series of lemmas which combine to yield a proof of Theorem 5. Suppose throughout that $\mathcal{F}^B$ is the restriction to B of $\mathcal{F}$, a p-forest labeled by S, for some $B \subseteq S$ with $|B| = b$ and $|S| = n$. To avoid trivialities, it is assumed throughout that $p_B > 0$. When convenient, as in the next lemma, it may be also be assumed (without loss of generality) that $S = [n] := \{1, \ldots, n\}$ and $B = [b]$ for some $b \in [n]$.

Lemma 8 *Conditionally given $|\mathcal{F}_i| = f_i$ for all $i \in [n]$, the random set $\mathcal{F}_1$ of children of 1 has uniform distribution over all subsets of size f_1 of $\{2, \ldots, n\}$, and for each $2 \leq i < n$ given also the subsets $\mathcal{F}_j$ of $[n]$ for all $j < i$, the random set $\mathcal{F}_i$ has uniform distribution over all subsets of size f_i of some subset of $[n]$ of size $n - 1 - f_1 - \cdots - f_{i-1}$, this subset of $[n]$ being determined by the $\mathcal{F}_j$ for $j < i$ and the constraint that $\mathcal{F}$ is a forest.*

Proof. This can be read from the proof of [15, Thm. 1.6].

Lemma 9 *For each $\mathbf{g} \in \mathbf{F}(B)$ and all vectors of non-negative counts $(f_i, i \in B)$ with $P(|\mathcal{F}_i| = f_i \text{ for all } i \in B) > 0$*

$$P(\mathcal{F}^B = \mathbf{g} \mid \ |\mathcal{F}_i| = f_i \text{ for all } i \in B) = \frac{(n - 1 - \sum_{i \in B} f_i)_{b - |\mathbf{g}| - 1}}{(n-1)_{b-1}} \prod_{i \in B} (f_i)_{|\mathbf{g}_i|}. \qquad (6)$$

Proof. The event $\mathcal{F}^B = \mathbf{g}$ is identical to the event that $\mathcal{F}_i \cap B = \mathbf{g}_i$ for all $i \in B$. For $B = [b] \subset S = [n]$, Lemma 8 shows that conditionally given $|\mathcal{F}_i| = f_i$ for all $i \in [b]$ there are

$$\prod_{m=1}^{b} \binom{n - 1 - \sum_{i=1}^{m-1} f_i}{f_m} = \frac{(n-1)!}{(n - 1 - \sum_{i=1}^{b} f_i)! \prod_{i=1}^{b} f_i!} \qquad (7)$$

equally likely possible choices of the sets $\mathcal{F}_i$ for $i \in [b]$. The number of these choices that make the event $(\mathcal{F}^B = \mathbf{g})$ occur is

$$\prod_{m=1}^{b} \binom{n-b-\sum_{i=1}^{m-1}(f_i - g_i)}{f_m - g_m} = \frac{(n-b)!}{(n-b-\sum_{i=1}^{b}(f_i-g_i))!\prod_{i=1}^{b}(f_i-g_i)!} \tag{8}$$

where $g_i := |\mathbf{g}_i|$, and the ratio of (8) to (7) simplifies to yield (6). To check the left-hand formula in (8), observe that given choices of the $\mathcal{F}_i$ have been made for $i < m$ in such a way that $|\mathcal{F}_i| = f_i$ and $\mathcal{F}_i \cap [b] = \mathbf{g}_i$ for all $i < m$, the choice of the set $\mathcal{F}_m$ of size f_m is subject firstly to the constraint that $\mathcal{F}$ is a forest, and secondly to the constraint that $\mathcal{F}_m \cap [b] = \mathbf{g}_m$. This means that there $f_m - g_m$ elements of $[n]-[b]$ to be chosen. The forest constraint forbids the choice of any of the $\sum_{i=1}^{m-1} f_i$ children of vertices $1, \ldots, m-1$ to be chosen. But due to previous choices, $\sum_{i=1}^{m-1} g_i$ of these forbidden vertices are contained in $[b]$, so there are exactly $\sum_{i=1}^{m-1}(f_i - g_i)$ forbidden vertices within $[n]-[b]$, and the $f_m - g_m$ vertices of $\mathcal{F}_m \cap ([n]-[b])$ are chosen from an allowed set of $n-b-\sum_{i=1}^{m-1}(f_i-g_i)$ vertices. Therefore, no matter what the $\mathcal{F}_i$ for $i < m$ such that $|\mathcal{F}_i| = f_i$ and $\mathcal{F}_i \cap [b] = \mathbf{g}_i$ for all $i < m$, the number of possible choices of $\mathcal{F}_m$ such that $\mathcal{F}_m \cap [b] = \mathbf{g}_m$ is the mth factor on the left side of (8). □

For the rest of this section let C_B denote the total number of children in $\mathcal{F}$ of all vertices in B:

$$C_B := |\mathcal{F} \cap (B \times S)| = \sum_{s \in B} |\mathcal{F}_s|.$$

Lemma 10 *For each $\mathbf{g} \in \mathbf{F}(B)$ with j tree components and each c with $P(C_B = c) > 0$,*

$$P(\mathcal{F}^B = \mathbf{g} \mid C_B = c) = \frac{(n-1-c)_{j-1}}{(n-1)_{b-1}} (c)_{b-j} \prod_{s \in B} \left(\frac{p_s}{p_B}\right)^{|\mathbf{g}_s|}. \tag{9}$$

Proof. Again, take $S = [n]$, $B = [b]$, and let $C_i := |\mathcal{F}_i|$ for $i \in [n]$. By application of (6),

$$P(\mathcal{F}^B = \mathbf{g} \mid C_B = c) = \frac{(n-1-c)_{j-1}}{(n-1)_{b-1}} E_c\left(\prod_{i=1}^{b} (C_i)_{|\mathbf{g}_i|}\right) \tag{10}$$

where E_c denotes expectation relative to the conditional distribution of $(C_1, \ldots, C_b)$ given $C_B = c$, which by Proposition 2 is a multinomial distribution with parameters c and $(p_1/p_B, \ldots, p_b/p_B)$. But this expectation can be evaluated by a calculation with the generating function of the multinomial distribution, and the result is (9). □

Recall that for $1 \le n \le N$ and $0 \le G \le N$ the *hypergeometic(n, N, G) distribution* is the distribution of the number of good elements that appear in a random subset of size n picked from a set of G good elements and $N - G$ bad elements [7].

Lemma 11
(i) *the distribution of C_B given $|\mathcal{F}| = m$ is* binomial (m, p_B)*;*
(ii) *given $|\mathcal{F}|$ and $C_B = c$, the distribution of $|\mathcal{F}^B|$ is* hypergeometic $(b-1, n-1, c)$.

Proof. Part (i) is immediate from Proposition 2. To obtain (ii), sum the expression (9) over all forests $\mathbf{g} \in \mathbf{F}(B)$ with ℓ edges and simplify using the multinomial expansion over forests (2) to see that

$$P(|\mathcal{F}^B| = \ell \,|\, C_B = c) = \frac{(n-1-c)_{b-\ell-1}(c)_\ell}{(n-1)_{b-1}}\binom{b-1}{\ell} = \binom{c}{\ell}\binom{n-1-c}{b-1-\ell}\binom{n-1}{b-1}^{-1}$$

which yields (ii). □

Proof of Theorem 5. Compare (9) and (3) to see that for each $c \in [n-1]$ the conditional distribution of $\mathcal{F}^B$ given $C_B = c$ is that of a $p(\cdot \,|\, B)$-forest, hence so is the unconditional distribution of $\mathcal{F}^B$. To compute the factorial moments of $|\mathcal{F}^B|$ recall that for indicator variables $X_i, i \in I$ and $r = 0, 1, 2, \ldots$ there is the formula

$$E\binom{\sum_{i \in I} X_i}{r} = \sum_{J \subseteq I : |J| = r} P(\cap_{j \in J}(X_j = 1)). \tag{11}$$

By standard applications of (11), for $S_{n,p}$ with binomial(n,p) distribution and $H_{n,N,G}$ with hypergeometric(n, N, G) distribution there are the formulae

$$E\binom{S_{n,p}}{r} = \binom{n}{r}p^r; \qquad E\binom{H_{n,N,G}}{r} = \binom{n}{r}\frac{(G)_r}{(N)_r}. \tag{12}$$

By application of these formulae and Lemma 11, for $\mathcal{F}$ with m edges the binomial moments of $|\mathcal{F}^B|$ are

$$E\binom{|\mathcal{F}^B|}{r} = E\left(E\left[\binom{|\mathcal{F}^B|}{r}\middle| C_B\right]\right) = \frac{(b-1)_r}{(n-1)_r}E\binom{C_B}{r} = \frac{(b-1)_r}{(n-1)_r}\binom{m}{r}p_B^r$$

and (4) follows. □

Examples. By application of (4) and (5), assuming that $\mathcal{F}$ has a fixed number k of tree components, so $\mu_r = (n-k)_r$, for each B with $|B| = b$ the restriction of $\mathcal{F}$ to B is a tree with probability

$$P(|\mathcal{F}^B| = b-1) = \frac{(n-k)_{b-1}}{(n-1)_{b-1}}\, p_B^{b-1}. \tag{13}$$

The restriction has two tree components with probability

$$P(|\mathcal{F}^B| = b-2) = (b-1)\left(\frac{(n-k)_{b-2}}{(n-1)_{b-2}}\, p_B^{b-2} - \frac{(n-k)_{b-1}}{(n-1)_{b-1}}\, p_B^{b-1}\right) \tag{14}$$

and so on. For p uniform, $p_B = b/n$, and the above probabilities have combinatorial interpretations as fractions of the total number $\binom{n-1}{k-1}n^{n-k}$ of forests of k rooted trees labeled by $[n]$. To illustrate with (13), the number of forests of k trees labeled by $[n]$ whose restriction to $[b]$ is a tree is

$$\frac{(n-k)_{b-1}}{(n-1)_{b-1}}\left(\frac{b}{n}\right)^{b-1}\binom{n-1}{k-1}n^{n-k}. \tag{15}$$

In particular, according to (15) for $k = 1$, there are $b^{b-1}n^{n-b}$ rooted trees labeled by $[n]$ whose restriction to $[b]$ is a tree. To check this, observe that such a tree is constructed by a unique sequence of choices according to the following three step procedure, where the numbers of choices in the first two steps are given by well known formulae of Cayley [5]:

1) pick an unrooted tree labeled by $[b]$, that is b^{b-2} possible choices;

2) pick a forest of b unrooted trees labeled by $[n]$, with one point of $[b]$ in each tree, that is bn^{n-b-1} choices,

3) let the set of edges of an unrooted tree labeled by $[n]$ be the union of the sets of edges of these $b+1$ trees, and pick a root from $[n]$, that is n choices.

The number of rooted trees labeled by $[n]$ whose restriction to $[b]$ is a tree is therefore

$$b^{b-2}\left(bn^{n-b-1}\right)n = b^{b-1}n^{n-b}. \tag{16}$$

For a survey of related enumerations see Moon [13].

3 The probability that $\mathcal{F}$ contains a particular set of edges.

For a random forest $\mathcal{F}$ labeled by S, and a set of edges $\mathbf{g} \subset S \times S$, it is a natural problem to calculate $P(\mathcal{F} \supseteq \mathbf{g})$, the probability that $\mathcal{F}$ contains each edge in the set $\mathbf{g}$. Obviously, this probability is zero unless $\mathbf{g}$ is a forest. Pemantle [14, Th. 4.2] found a determinant formula for probabilities of this kind derived from a uniform random spanning tree of a graph. In the model of random forests considered here, there is the following simpler result:

Theorem 12 *Suppose that $\mathcal{F}$ is a p-forest labeled by S with $|S| = n$. Then for each rooted forest $\mathbf{g}$ labeled by S with r edges*

$$P(\mathcal{F} \supseteq \mathbf{g}) = \frac{E(|\mathcal{F}|)_r}{(n-1)_r} \prod_{s \in S} p_s^{|\mathbf{g}_s|}. \tag{17}$$

To illustrate this formula, for any two distinct s and s' in S, the probability that $\mathcal{F}$ contains a particular edge (s, s') is

$$P(s \xrightarrow{\mathcal{F}} s') = \frac{E|\mathcal{F}|}{(n-1)} p_s \tag{18}$$

and for distinct t and t' in S, with $(s, s') \neq (t', t)$ and $s' \neq t'$, the probability that $\mathcal{F}$ contains both (s, s') and (t, t') is

$$P((s \xrightarrow{\mathcal{F}} s') \cap (t \xrightarrow{\mathcal{F}} t')) = \frac{E(|\mathcal{F}|(|\mathcal{F}|-1))}{(n-1)(n-2)} p_s p_t. \tag{19}$$

In particular, for such (s, s') and (t, t') the events $(s \xrightarrow{\mathcal{F}} s')$ and $(t \xrightarrow{\mathcal{F}} t')$ are independent if $\mathcal{F}$ is a p-tree, and negatively correlated if $\mathcal{F}$ is a p-forest of k trees for $k \geq 2$.

Proof of Theorem 12. By conditioning on $|\mathcal{F}|$ it is enough to consider the case when $\mathcal{F}$ has a fixed number m of edges. The left side of (17) in this case is a sum over all forests $\mathbf{f} \supseteq \mathbf{g}$ of $P(\mathcal{F} = \mathbf{f})$ defined by the product formula (1). Thus (17) can be read from the following lemma, where the probabilities p_s are replaced by variables x_s not subject to the constraints of a probability distribution:

Lemma 13 *For each rooted forest $\mathbf{g}$ labeled by S and each integer $m \geq |\mathbf{g}|$*

$$\sum_{\mathbf{f}:|\mathbf{f}|=m, \mathbf{f} \supseteq \mathbf{g}} \prod_{s \in S} x_s^{|\mathbf{f}_s|} = \binom{|S|-1-|\mathbf{g}|}{m-|\mathbf{g}|} \left(\prod_{s \in S} x_s^{|\mathbf{g}_s|} \right) \left(\sum_{s \in S} x_s \right)^{m-|\mathbf{g}|}. \tag{20}$$

where the sum on the left is over all rooted forests $\mathbf{f}$ labeled by S with m edges containing $\mathbf{g}$.

Proof. It is enough to consider $S = [n]$. Let $|\mathbf{g}_i| = g_i$. By a reprise of the argument which yielded (8), the number of forests $\mathbf{f}$ labeled by $[n]$ such that $\mathbf{f}$ contains $\mathbf{g}$ and $|\mathbf{f}_i| = f_i$ for all $i \in [n]$ is

$$\prod_{j=1}^{n} \binom{m-1-\sum_{i=1}^{j-1}(f_i - g_i)}{f_j - g_j} = \binom{n-1-|\mathbf{g}|}{m-|\mathbf{g}|}\binom{m-|\mathbf{g}|}{f_1-g_1,\ldots,f_n-g_n}$$

which gives the identity of coefficients of $\prod_{s\in[n]} x_s^{f_s}$ in (20). □

Examples. The special case of (20) when $\mathbf{g}$ is the trivial forest with no edges is the basic multinomial expansion over forests (2). Take the $x_s \equiv 1$ in (20) to deduce that for every rooted forest $\mathbf{g}$ labeled by $[n]$ with j tree components, and every $1 \le k \le j$, the number of rooted forests $\mathbf{f}$ labeled by $[n]$ which contain $\mathbf{g}$ and have k tree components is $\binom{j-1}{k-1} n^{j-k}$. For another proof of this enumeration, and various applications, see [15].

Alternative proof of (4). Since

$$|\mathcal{F}^B| = \sum_{(s,t)\in B\times B} 1(s \xrightarrow{\mathcal{F}} t) \tag{21}$$

the general formula (11) gives for $r = 1, 2, \ldots, b-1$

$$E\binom{|\mathcal{F}^B|}{r} = \sum_{\mathbf{g}\subseteq B\times B:|\mathbf{g}|=r} P(\mathcal{F} \supseteq \mathbf{g}) \tag{22}$$

The probability $P(\mathcal{F} \supseteq \mathbf{g})$ is zero unless $\mathbf{g}$ is a rooted forest with r edges, and for such $\mathbf{g}$ this probability is evaluated by Theorem 12. Thus

$$E\binom{|\mathcal{F}^B|}{r} = \sum_{\mathbf{g}\in\mathbf{F}(B):|\mathbf{g}|=r} \frac{E(|\mathcal{F}|)_r}{(n-1)_r} \prod_{s\in B} p_s^{|\mathbf{g}_s|} = \frac{E(|\mathcal{F}|)_r}{(n-1)_r}\binom{b-1}{r} p_B^r \tag{23}$$

where the second equality is due to (2). □

4 Random thinning of edges

There is one case where a substantial simplification occurs in the formulae (4) and (5). Suppose that $|\mathcal{F}|$ has a binomial distribution with parameters m and q for some $q \in [0,1]$.

Then from (12) and (4), the distribution of $|\mathcal{F}^B|$ has rth factorial moment

$$E(|\mathcal{F}^B|)_r = \frac{(b-1)_r}{(n-1)_r}(m)_r q^r p_B^r. \tag{24}$$

If $m = n-1$ this expression simplifies to $(b-1)_r(qp_B)^r$, which is the rth factorial moment of the binomial distribution with parameters $b-1$ and qp_B. This yields part (i) of the following following corollary of Theorem 5. Both parts follow easily from Lemma 11.

Corollary 14 *Suppose $\mathcal{F}$ is a p-forest labeled by S with $|S| = n$, and that the number of edges of $\mathcal{F}$ has binomial$(n-1, q)$ distribution for some $q \in [0,1]$. Then for each $B \subseteq S$ with $|B| = b$,*
(i) *the restricted forest $\mathcal{F}^B$ is a $p(\cdot \mid B)$-forest whose number of edges $|\mathcal{F}^B|$ has binomial$(b-1, qp_B)$ distribution.*
(ii) *the number $|\mathcal{F}^B|$ of edges of $\mathcal{F}$ in $B \times B$, and the number of edges of $\mathcal{F}$ in $B \times B^c$ are independent, and the latter number has binomial$(n-b, qp_B)$ distribution.*

Let $\mathcal{T}$ be a p-tree labeled by S, and let $\mathcal{F}$ be derived from $\mathcal{T}$ by retaining each of the $n-1$ edges of $\mathcal{T}$ independently with probability q. Call $\mathcal{F}$ a *q-thinning* of $\mathcal{T}$. By application of Proposition 3, $\mathcal{F}$ is a p-forest, and $|\mathcal{F}|$ has the binomial$(n-1, q)$ distribution supposed in the above corollary. To restate the corollary, *the restriction to B of a q-thinning of a p-tree has the same distribution as a qp_B-thinning of a $p(\cdot \mid B)$-tree.* Even for p uniform and $q = 1$ this result does not seem evident without calculation. Neither does the independence property (ii) seem obvious even in this case.

5 A moment identity.

In the setting of Lemma 11, there is the following expression for the distribution of $|\mathcal{F}^B|$:

$$P(|\mathcal{F}^B| = \ell) = \binom{n-1}{b-1}^{-1} E\left[\binom{n-1-C_B}{b-\ell-1}\binom{C_B}{\ell}\right] \tag{25}$$

where C_B has binomial (m, p_B) distribution given that $|\mathcal{F}| = m$. Compare (25), (5) and (12) to see that the following moment identity (26) must hold for a binomially distributed random variable Y, with some restrictions on x. But then the identity must hold as stated, by straightforward extrapolations. As a check, the alternate proof given below reduces the moment identity to a known identity for binomial coefficients.

Lemma 15 *Let Y be a random variable with all moments finite. Then for all real x and all non-negative integers a and b*

$$E\left[\binom{x-Y}{a}\binom{Y}{b}\right] = \sum_{j=0}^{a}(-1)^j\binom{b+j}{j}\binom{x-b-j}{a-j}E\binom{Y}{b+j} \tag{26}$$

Proof. By linearity of the expectation operator E, it suffices to prove the formula for a constant random variable Y, say $Y = y$ for some real y. Then the formula reduces easily to

$$\binom{x-y}{a} = \sum_{j=0}^{a}(-1)^j\binom{x-b-j}{a-j}\binom{y-b}{j}. \tag{27}$$

Replace $x-b$ by x and $y-b$ by $-z$ to see that this amounts to

$$\binom{x+z}{a} = \sum_{j=0}^{a}\binom{x-j}{a-j}\binom{z+j-1}{j} \tag{28}$$

for all real x and z, which is a known identity for binomial coefficients (replace n by a, x by $z-1$ and y by $x-a$ in Gould [8][(3.2)]).

References

[1] D. Aldous and J. Pitman. The entrance boundary of the additive coalescent. Paper in preparation, 1997.

[2] D.J. Aldous. The continuum random tree II: an overview. In M.T. Barlow and N.H. Bingham, editors, *Stochastic Analysis*, pages 23–70. Cambridge University Press, 1991.

[3] B. Bollobás. *Random Graphs*. Academic Press, London, 1985.

[4] Y. D. Burtin. On a simple formula for random mappings and its applications. *J. Appl. Probab.*, 17:403 – 414, 1980.

[5] A. Cayley. A theorem on trees. *Quarterly Journal of Pure and Applied Mathematics*, 23:376–378, 1889. (Also in *The Collected Mathematical Papers of Arthur Cayley. Vol XIII*, 26-28, Cambridge University Press, 1897).

[6] S.N. Evans and J. Pitman. Construction of Markovian coalescents. Technical Report 465, Dept. Statistics, U.C. Berkeley, 1996. Revised May 1997. To appear in *Ann. Inst. Henri Poincaré.*

[7] W. Feller. *An Introduction to Probability Theory and its Applications,* Vol 1,3rd ed. Wiley, New York, 1968.

[8] H.W. Gould. *Combinatorial Identities: A standardized set of tables listing 500 binomial coefficient summations.* West Virginia University, Morgantown, W. Va., 1972.

[9] A. Hurwitz. Über Abel's Verallgemeinerung der binomischen Formel. *Acta Math.*, 26:199–203, 1902.

[10] J.Jaworski. On a random mapping (T, P_j). *J. Appl. Probab.*, 21:186 – 191, 1984.

[11] T. Łuczak and B. Pittel. Components of random forests. *Combinatorics, Probability and Computing*, 1:35–52, 1992.

[12] J.A. Mann, B. Pittel, and W. A. Woycznski. Random tree-type partitions as a model for acyclic polymerization: Holtsmark (3/2 stable) distribution of the supercritical gel. *Ann. Probab.*, 18:319–341, 1990.

[13] J.W. Moon. *Counting Labelled Trees.* Canadian Mathematical Congress, 1970. Canadian Mathematical Monographs No. 1.

[14] R. Pemantle. Uniform random spanning trees. In J. Laurie Snell, editor, *Topics in Contemporary Probability*, pages 1–54, Boca Raton, FL, 1995. CRC Press.

[15] J. Pitman. Coalescent random forests. Technical Report 457, Dept. Statistics, U.C. Berkeley, 1996. To appear in *J. Comb. Theory A*. Available via http://www.stat.berkeley.edu/users/pitman.

[16] J. Pitman. Abel-Cayley-Hurwitz multinomial expansions associated with random mappings, forests and subsets. Technical Report 498, Dept. Statistics, U.C. Berkeley, 1997. Available via http://www.stat.berkeley.edu/users/pitman.

[17] J. Pitman. The asymptotic behavior of the Hurwitz binomial distribution. Technical Report 500, Dept. Statistics, U.C. Berkeley, 1997.

[18] J. Pitman. Enumerations of trees and forests related to branching processes and random walks. In D. Aldous and J. Propp, editors, *Microsurveys in Discrete Probability*, number 41 in DIMACS Ser. Discrete Math. Theoret. Comp. Sci, pages 163–180, Providence RI, 1998. Amer. Math. Soc.

[19] A. Rényi. On the enumeration of trees. In R. Guy, H. Hanani, N. Sauer, and J. Schonheim, editors, *Combinatorial Structures and their Applications*, pages 355–360. Gordon and Breach, New York, 1970.

[20] S. M. Ross. A random graph. *J. Appl. Probab.*, 18:309–315, 1981.

[21] R. Stanley. Enumerative combinatorics, vol. 2. Book in preparation, to be published by Cambridge University Press, 1996.