# Coalescent Random Forests*

by Jim Pitman

Technical Report No. 457
Department of Statistics
University of California
367 Evans Hall # 3860
Berkeley, CA 94720-3860

Revised version.
Accepted for publication in
J. Combinatorial Theory A
as of July 2,1998

*Research supported in part by N.S.F. Grants MCS94-04345 and DMS 97-03961

**Abstract**

Various enumerations of labeled trees and forests, including Cayley's formula $n^{n-2}$ for the number of trees labeled by $[n]$, and Cayley's multinomial expansion over trees, are derived from the following *coalescent construction* of a sequence of random forests $(R_n, R_{n-1}, \ldots, R_1)$ such that $R_k$ has uniform distribution over the set of all forests of $k$ rooted trees labeled by $[n]$. Let $R_n$ be the trivial forest with $n$ root vertices and no edges. For $n \geq k \geq 2$, given that $R_n, \ldots, R_k$ have been defined so that $R_k$ is a rooted forest of $k$ trees, define $R_{k-1}$ by addition to $R_k$ of a single edge picked uniformly at random from the set of $n(k-1)$ edges which when added to $R_k$ yield a rooted forest of $k-1$ trees. This coalescent construction is related to a model for a physical process of clustering or coagulation, the *additive coalescent* in which a system of masses is subject to binary coalescent collisions, with each pair of masses of magnitudes $x$ and $y$ running a risk at rate $x+y$ of a coalescent collision resulting in a mass of magnitude $x+y$. The transition semigroup of the additive coalescent is shown to involve probability distributions associated with a multinomial expansion over rooted forests.

# 1 Introduction

Let $\mathcal{T}_n$ denote the set of all trees labeled by $[n] := \{1, \ldots, n\}$. Cayley's [14] formula $\#\mathcal{T}_n = n^{n-2}$ is a well known consequence of the bijection between $\mathcal{T}_n$ and $[n]^{n-2}$ set up by Prüfer's [51] coding of trees. See [19, 37, 38, 60] for background, alternative proofs of Cayley's formula, and related codings and enumerations. One purpose of this paper is to show how various enumerations of labeled trees and forests, including Cayley's formula, follow easily from a very different construction of random forests by a *coalescent process*. A second purpose is to relate this construction to various models of coalescent processes which have found applications in statistical physics and polymer chemistry [34, 32, 31, 65, 22, 12, 8], computer science [64, 26], genetics [25], combinatorics [13, 3, 7], and astronomy [59]. A third purpose is to lay combinatorial foundations for the study undertaken in companion papers [9, 18] of asymptotic properties of the *additive coalescent* process, in which a system of masses is subject to binary coalescent collisions, with each pair of masses of magnitudes $x$ and $y$ running a risk at rate $x+y$ of a coalescent collision resulting in a mass of magnitude $x+y$. These asymptotics, which allow the definition of the additive coalescent process to be extended to an infinite number of masses, are related to the lengths of excursion intervals of a Brownian motion [49] and Aldous's concept of a *continuum random tree* associated with a Brownian excursion [4, 5, 6].

The paper is organized as follows. Section 2 derives some basic enumerations for labeled trees and forests by a combinatorial version of the coalescent construction. Section 3 interprets these enumerations probabilistically by construction of a uniformly distributed random tree as the last term of a coalescent sequence of random forests. Section 4 relates this construction to various known results concerning random partitions derived from coalescent processes and models for random forests. Section 4.4 indicates some applications to random graphs. Section 5 shows how Cayley's multinomial expansion over trees can be deduced from the basic coalescent construction, and offers some variations and probabilistic interpretations of this multinomial expansion. Section 6 shows how an additive coalescent process with arbitrary initial condition can be derived from a coalescent construction of random forests, and deduces a formula for the transition semigroup of the additive coalescent which is related to a multinomial expansion over rooted forests.

# 2 Basic enumerations

Except when specified otherwise, a tree $t$ is assumed to be unrooted, and labeled by some finite set $S$. Then call $t$ a *tree over* $S$. Write $\#S$ for the number of elements of $S$. Call a two element subset $\{a, b\}$ of $S$, which may be denoted instead $a \leftrightarrow b$, an *edge* or a *bond*. A tree $t$ over $S$ is identified by its set of $\#S - 1$ edges. A *forest over* $[n]$ is a graph with vertex set $[n]$ whose connected components form a collection of trees labeled by the sets of some partition of $[n]$. Note that each forest over $[n]$ with $k$ tree components has $n - k$ edges.

## 2.1 Rooted forests

A *rooted forest over* $[n]$ is a forest labeled by $[n]$ together with a choice of a root vertex for each tree in the forest. Let $\mathcal{R}_{k,n}$ be the set of all rooted forests of $k$ trees over $[n]$. A rooted forest is identified by its *digraph*, that is its set of directed edges, sometimes denoted $a \to b$ instead of $(a, b)$, with edges directed away from the roots. Say one rooted forest $r$ *contains* another rooted forest $s$ if the digraph of $r$ contains the digraph of $s$. Call a sequence of rooted forests $(r_i)$ *refining* if $r_i$ contains $r_j$ for $i < j$. The following lemma is the simpler equivalent for rooted forests of an enumeration of unrooted forests due to Moon [35], which appears later as Lemma 3.

**Lemma 1** *For each forest $r_k$ of $k$ rooted trees over $[n]$, the number of rooted trees over $[n]$ that contain $r_k$ is $n^{k-1}$.*

**Proof.** For $r_k \in \mathcal{R}_{k,n}$ let $N(r_k)$ denote the number of rooted trees over $[n]$ that contain $r_k$, and let $N^*(r_k)$ denote the number of refining sequences $(r_1, r_2, \ldots, r_k)$ with $r_j \in \mathcal{R}_{jn}$ for $1 \le j \le k$. Any tree $r_1$ which contains $r_k$ has $(n-1)-(n-k) = k-1$ bonds more than $r_k$. So to choose a refining sequence $(r_1, r_2, \ldots, r_k)$ starting from any particular $r_1$ that contains $r_k$, there are $k-1$ bonds of $r_1$ that could be deleted to choose $r_2$, then $k-2$ bonds of $r_2$ that could be deleted to choose $r_3$, and so on. Therefore

$$N^*(r_k) = N(r_k)(k-1)! \tag{1}$$

Now consider choosing such a refining sequence $(r_1, r_2, \ldots, r_k)$ in reverse order. Given the choice of $(r_k, r_{k-1}, \ldots, r_j)$ for some $k \ge j \ge 2$ the number of possible choices of $r_{j-1}$ is the number of ways to choose the directed edge $a \to b$ that is in $r_{j-1}$ but not $r_j$. But $a$ can be any vertex in $[n]$, and then $b$ any one of the $j-1$ roots of the $j-1$ trees in $r_j$ that do not contain $a$. (If $b$ is not one of those roots then the resulting digraph is not a rooted forest.) So the number of possible choices of $r_{j-1}$ given $(r_k, \ldots, r_j)$ is always $n(j-1)$. This yields

$$N^*(r_k) = n^{k-1}(k-1)! \tag{2}$$

Now (1) and (2) imply $N(r_k) = n^{k-1}$. □

For $k = n$ Lemma 1 shows that the number of rooted trees over $[n]$ is $\#\mathcal{R}_{1n} = n^{n-1}$, which is equivalent to Cayley's formula $\#\mathcal{T}_n = n^{n-2}$. Formula (2) for $k = n$ gives

$$\#\{\text{ refining } (r_1, r_2, \ldots, r_n) : r_j \in \mathcal{R}_{jn} \text{ for } 1 \le j \le n\} = n^{n-1}(n-1)! \tag{3}$$

For $r_k \in \mathcal{R}_{k,n}$ let $N^{**}(r_k)$ denote the number of these refining sequences of rooted forests with the $k$th term specified equal to $r_k$. This is the number of ways to choose $(r_1, \ldots, r_{k-1})$ times the number of ways to choose $(r_{k+1}, \ldots, r_n)$, that is from (2)

$$N^{**}(r_k) = N^*(r_k)\,(n-k)! = n^{k-1}\,(k-1)!\,(n-k)! \tag{4}$$

Because this number does not depend on the choice of $r_k \in \mathcal{R}_{k,n}$, dividing (3) by (4) yields the number of rooted forests of $k$ trees over $[n]$:

$$\#\mathcal{R}_{k,n} = \frac{n^{n-1}(n-1)!}{n^{k-1}(k-1)!(n-k)!} = \binom{n}{k} k n^{n-k-1} = \binom{n-1}{k-1} n^{n-k} \tag{5}$$

This enumeration appears as (8a) in Riordan [55] with a proof using generating functions and the Lagrange inversion formula. Let $\mathcal{R}^0_{k,n}$ denote the subset of $\mathcal{R}_{k,n}$ consisting of all

rooted forests over $[n]$ whose set of roots is $[k]$. An $r_k \in \mathcal{R}_{k,n}$ is specified by first picking its set of $k$ roots, then picking a forest with those roots. So (5) amounts to the following result stated by Cayley [14] and proved by Rényi [52]:

$$\#\mathcal{R}^0_{k,n} = kn^{n-k-1} \tag{6}$$

For alternative proofs and equivalents of this formula see [38, 39, 63] and [10, Lemma 17]. The same method yields easily the following result, which includes both Lemma 1 and formula (5) as special cases:

**Proposition 2** *For each $1 \le k \le j \le n$, and each forest $r_j$ of $j$ rooted trees over $[n]$, the number of forests of $k$ rooted trees over $[n]$ that contain $r_j$ is $\binom{j-1}{k-1} n^{j-k}$.*

## 2.2 Unrooted forests

Let $\mathcal{F}_{k,n}$ be the set of unrooted forests of $k$ trees over $[n]$. The analog of Lemma 1 for unrooted forests is more complicated:

**Lemma 3** (Moon [35]): *If $f_k \in \mathcal{F}_{k,n}$ consists of $k$ trees of sizes $n_1, \ldots, n_k$, where $\sum_i n_i = n$, then the number $N(f_k)$ of trees $t \in \mathcal{T}_n$ which contain $f_k$ is*

$$N(f_k) = \left( \prod_{i=1}^{k} n_i \right) n^{k-2} \tag{7}$$

**Proof.** The number of rooted trees over $[n]$ whose edge set (with directions ignored) contains $f_k$ is

$$nN(f_k) = \left( \prod_{i=1}^{k} n_i \right) n^{k-1}$$

where the left-hand evaluation is obvious, and the right-hand evaluation is obtained by first choosing roots for the $k$ tree components of $f_k$, and then applying Lemma 1. □

See Stanley ([62], Exercise 2.11) for a generalization of Lemma 3 which can be obtained by the same method, and an application to enumeration of spanning trees of a general graph. Section 5 shows how Moon's derivation of Lemma 3 can be reversed to deduce Cayley's multinomial expansion over trees.

# 3 Random forests

The following two theorems are probabilistic expressions of the enumeration of refining sequences of rooted forests described in Section 2.1.

**Theorem 4** *The following three descriptions* (i), (ii) *and* (iii), *of the distribution of random sequence* $(R_1, R_2, \ldots, R_n)$ *of rooted forests over* $[n]$, *are equivalent, and imply that*

$$R_k \text{ has uniform distribution over } \mathcal{R}_{k,n} \text{ for each } 1 \le k \le n. \tag{8}$$

(i) $R_1$ *is a uniformly distributed rooted tree over* $[n]$, *and given* $R_1$, *for each* $1 \le k \le n$ *the forest* $R_k$ *is derived by deletion from* $R_1$ *of* $k-1$ *edges* $e_j, 1 \le j \le k-1$, *where* $(e_j, 1 \le j \le n-1)$ *is a uniformly distributed random permutation of the set of* $n-1$ *edges of* $R_1$*;*

(ii) $R_n$ *is the trivial digraph, and for* $n \ge k \ge 2$, *given* $R_n, \ldots, R_k$ *with* $R_k \in \mathcal{R}_{k,n}$, *the forest* $R_{k-1} \in \mathcal{R}_{k-1,n}$ *is derived from* $R_k$ *by addition of a single directed edge picked uniformly at random from the set of* $n(k-1)$ *directed edges which when added to* $R_k$ *yield a rooted forest of* $k-1$ *trees over* $[n]$.

(iii) *the sequence* $(R_1, R_2, \ldots, R_n)$ *has uniform distribution over the set of all* $(n-1)!n^{n-1}$ *refining sequences of rooted forests* $(r_1, r_2, \ldots, r_n)$ *with* $r_k \in \mathcal{R}_{k,n}$ *for each* $1 \le k \le n$.

**Proof.** The equivalence of (i), (ii), and (iii) is evident from the enumeration (3), and the uniform distribution of $R_k$ follows from (4). □

The next theorem is just a reformulation of Theorem 4 in terms of unrooted forests instead of rooted forests. While the correspondence between (i) and (i)' and between (iii) and (iii)' in the two formulations is obvious, in the unrooted formulation the distributions of the intermediate random forests displayed in (9) are not uniform, and the coalescent description (ii)' is consequently more complicated. The rule (10) in (ii)' for picking which pair of trees to join in the coalescent process obtained by time-reversal of (i)' appears without proof in Yao [64, Lemma 2].

**Theorem 5** *The following three descriptions* (i)',(ii)' *and* (iii)', *for the distribution of sequence* $(F_1, F_2, \ldots, F_n)$ *of random forests over* $[n]$, *are equivalent, and imply that for each* $1 \le k \le n$ *and each forest* $f_k \in \mathcal{F}_{k,n}$ *comprising* $k$ *trees with sizes* $n_1, \ldots, n_k$ *in some arbitrary order,*

$$P(F_k = f_k) = \frac{\prod_{i=1}^{k} n_i}{n^{n-k}\binom{n-1}{k-1}} \tag{9}$$

(i)' *$F_1$ is a uniform random tree over $[n]$, and given $F_1$, for each $1 \le k \le n$ the forest $F_k$ is derived by deletion from $F_1$ of $k-1$ edges $e_j, 1 \le j \le k-1$, where $(e_j, 1 \le j \le n-1)$ is a uniform random permutation of the set of $n-1$ edges of $F_1$;*

(ii)' *$F_n$ is the trivial forest, with $n$ vertices and no edges, and for $n \ge k \ge 2$, given $F_n, \ldots, F_k$ where $F_k$ is a forest over $[n]$ with $k$ tree components, say $\{T_1, \ldots, T_k\}$ where $T_i$ is a set of size $n_i$ and $\sum_{i=1}^k n_i = n$, the forest $F_{k-1} \in \mathcal{F}_{k-1,n}$ is derived from $F_k$ by addition of a single edge $a \leftrightarrow b$ according to the following rule: first*

$$\text{pick } (i,j) \text{ with probability } \frac{n_i + n_j}{n(k-1)} \text{ for } 1 \le i < j \le k, \tag{10}$$

*then pick $a$ and $b$ independently and uniformly at random from $T_i$ and $T_j$ respectively.*

(iii)' *the sequence $(F_1, F_2, \ldots, F_n)$ has uniform distribution over the set of all $(n-1)!n^{n-2}$ refining sequences of forests $(f_1, f_2, \ldots, f_n)$ such that $f_k \in \mathcal{F}_{k,n}$ for every $1 \le k \le n-1$.*

**Proof.** The equivalence of descriptions (i)' and (iii)' is obvious from Cayley's formula $\#\mathcal{T}_n = n^{n-2}$. From either of these descriptions, the forest $F_k$ is determined by a *choice* of a tree $t \in \mathcal{T}_n$ and a subset of $k-1$ bonds of $t$, and there are $n^{n-2}\binom{n-1}{k-1}$ equally likely choices. The number of such choices which make $F_k = f_k$ is the number of trees $t \in \mathcal{T}_n$ which contain $f_k$, as displayed in (7). The ratio of these two numbers yields the probability (9).

To check that the description (ii)' is equivalent to (i)' and (iii)' it suffices to show that (ii)' holds for $(F_1, \ldots, F_n)$ defined by unrooting a sequence $(R_1, \ldots, R_n)$ satisfying the conditions of Theorem 4. This can be verified as follows using the conditional distribution of $R_{k-1}$ given $R_k$ described in condition (ii) of Theorem 4. Consider the conditional probability that $F_{k-1} = f_{k-1}$ given $F_k = f_k$ where $f_{k-1}$ is obtained by adding a single edge $a \leftrightarrow b$ to $f_k$, where $a \in T_i$ and $b \in T_j$ for some $1 \le i < j \le k$. In terms of $R_k$ and $R_{k-1}$, this edge $a \leftrightarrow b$ is added iff

*either* vertex $a$ is the root of $T_i$ in $R_k$ and $R_{k-1}$ adds the directed edge $b \to a$ to $R_k$, which happens with probability $\frac{1}{n_i} \times \frac{1}{n(k-1)}$

*or* vertex $b$ is the root of $T_j$ in $R_k$ and $R_{k-1}$ adds the directed edge $a \to b$ to $R_k$, which happens with probability $\frac{1}{n_j} \times \frac{1}{n(k-1)}$

So the conditional probability that $F_{k-1} = f_{k-1}$ given $F_k = f_k$ is

$$\left(\frac{1}{n_i} + \frac{1}{n_j}\right) \frac{1}{n(k-1)} = \frac{n_i + n_j}{n(k-1)} \left(\frac{1}{n_i}\right) \left(\frac{1}{n_j}\right) \tag{11}$$

Because the sequence $(F_1, \ldots, F_n)$ has the Markov property, so does the reversed sequence $(F_n, \ldots, F_1)$. The expression (11) therefore gives the conditional probability that $F_{k-1} = f_{k-1}$ given $(F_n, \ldots, F_k)$ with $F_k = f_k$, which is condition (ii)'. □

**Alternative derivation of (11).** As a check, the conditional probability (11) can also be derived as follows [59]. Use Bayes' rule $P(A|B) = P(AB)/P(B)$ to compute

$$P(F_{k-1} = f_{k-1}|F_k = f_k) = \frac{P(F_{k-1} = f_{k-1})}{P(F_k = f_k)} P(F_k = f_k|F_{k-1} = f_{k-1})$$

and use (9) and description (i)' to evaluate the right side. This gives

$$P(F_{k-1} = f_{k-1}|F_k = f_k) = \frac{n^{n-k}\binom{n-1}{k-1}(n_i + n_j)\prod_{\ell \notin \{i,j\}} n_\ell}{n^{n-(k-1)}\binom{n-1}{k-2}\prod_\ell n_\ell} \frac{1}{n-k+1}$$

Since $\binom{n-1}{k-1}/\binom{n-1}{k-2} = (n-k+1)/(k-1)$ this expression reduces to (11).

## 3.1 Related coalescent constructions.

Motivated by an application to the theory of random graphs indicated in section 4.4, Aldous [3] and Buffet-Pulé [13] considered the following coalescent construction of a random element $F_1$ of $\mathcal{T}_n$, which is similar but not equivalent to the above construction (ii)':

(ii)" *Let $F_n$ be the trivial graph with no edges, and given that $F_n, \ldots, F_k$ have been defined with $F_k \in \mathcal{F}_{k,n}$, let $F_{k-1} \in \mathcal{F}_{k-1,n}$ be obtained by adding to $F_k$ a single edge picked uniformly at random from the set of all edges which when added to $F_k$ yield a forest of $k-1$ trees over $[n]$.*

As noted by Aldous, for $n \geq 4$ the final random tree $F_1$ generated by (ii)" does not have a uniform distribution, though Aldous conjectures that the asymptotic behaviour for large $n$ of some features of this random tree are similar to asymptotics of uniform random elements of $\mathcal{T}_n$ surveyed in [4, 5, 6].

A natural generalization of both coalescent constructions (ii)' and (ii)" can be made as follows, in terms of a discrete-time forest-valued Markov chain. Similar Markov chains with a continuous time parameter have been studied in the physical science literature [34, 32, 31, 22, 12] as models for processes of polymerization and coagulation. Let $\kappa(x, y)$ be a positive symmetric function of pairs of positive integers $(x, y)$, called a *collision rate*

*kernel*. As before, let $F_n$ be the trivial graph with no edges, and grow random forests $F_n, F_{n-1} \ldots, F_1$, with $F_k \in \mathcal{F}_{k,n}$, as follows:

*Given that $F_n, \ldots, F_k$ have been defined with $F_k$ a forest over $[n]$ of $k$ trees of sizes $n_1, \ldots, n_k$, let $F_{k-1}$ be derived from $F_k$ by adding an edge joining vertices picked independently and uniformly at random from the $i$th and $j$th of these trees, where $(i,j)$ is picked with probability proportional to $\kappa(n_i, n_j)$ for $1 \leq i < j \leq k$.*

Call such a sequence of random forests, with state space the set $\mathcal{F}_n := \cup_{k=1}^n \mathcal{F}_{k,n}$ of unrooted forests over $[n]$, starting with $F_n$ the trivial forest with $n$ singleton components and no edges, and terminating with a random tree $F_1$, a *discrete-time $\mathcal{F}_n$-valued $\kappa$-coalescent*. This process is a Markov chain with state space $\mathcal{F}_n$ whose transition probabilities are determined by the collision rate kernel $\kappa$. It is easily seen that the construction (ii)' gives the *discrete-time $\mathcal{F}_n$-valued additive coalescent* with kernel $\kappa(x,y) = x + y$, while Aldous's model (ii)" is the *discrete-time $\mathcal{F}_n$-valued multiplicative coalescent* with kernel $\kappa(x,y) = xy$. Here $\kappa(x,y)$ may be interpreted as a collision rate between trees of size $x$ and size $y$ in a continuous time Markov chain with state space $\mathcal{F}_n$. The discrete time $\mathcal{F}_n$-valued $\kappa$-coalescent is then the embedded discrete time chain defined by the sequence of distinct states of the continuous time coalescent.

Other models for random forests have been studied, including dynamic models featuring a stochastic equilibrium between processes of addition and deletion of edges. See [30, §7] for a brief survey of the literature of these models.

# 4 Random partitions

In applications of coalescent processes, the distribution of sizes of various clumps is of primary importance. The state of a coalescent process of $n$ particles is often regarded as a *partition of $n$*, that is a unordered collection of positive integer *clump sizes* with sum $n$. A partition of $n$ may be denoted $1^{m_1} 2^{m_2} \cdots n^{m_n}$ to indicate that there are $m_i$ clumps of size $i$ for each $1 \leq i \leq n$, where $n = \sum_i i m_i$ and the number of clumps is $\sum_i m_i$. If clumps are regarded as sets of labeled particles, the state of a coalescent process may be represented as a partition of the set $[n]$. In the random forest models of the previous section, each clump of particles also has an internal tree structure. Such models arise naturally in polymer chemistry [65], but the internal tree stucture may be ignored in other settings. Let

$$\mathcal{F}_n := \cup_{k=1}^n \mathcal{F}_{k,n}, \text{ the set of all forests over } [n]$$

$$\mathcal{P}_{[n]} := \text{ the set of all partitions of } [n]$$

$$\mathcal{P}_n := \text{ the set of all partitions of } n$$

There are natural projections from $\mathcal{F}_n$ onto $\mathcal{P}_{[n]}$ onto $\mathcal{P}_n$, say $f \to \Pi \to \pi$, where $\Pi$ is the partition of $[n]$ generated by the tree components of $f$, and $\pi$ is the partition of $n$ generated by the sizes of components of $\Pi$. By use of these projections and the standard criterion for a function of a Markov chain to be Markov, the discrete time $\kappa$-coalescent process defined in the previous section as a Markov chain with statespace $\mathcal{F}_n$ induces corresponding Markov chains with state space $\mathcal{P}_{[n]}$ and $\mathcal{P}_n$. In particular, the additive coalescent with kernel $\kappa(x,y) = x + y$ will be viewed in this section as a $\mathcal{P}_{[n]}$-valued process. A discrete time $\kappa$-coalescent process with state space $\mathcal{P}_{[n]}$ is a Markovian sequence $(\Pi_n, \ldots, \Pi_1)$ of coarsening random partitions of $[n]$, starting with $\Pi_n = \{\{1\}, \{2\}, \ldots, \{n\}\}$ and terminating with $\Pi_1 = \{\{1, 2, \ldots n\}\}$, such that $\Pi_k$ is a partition of $[n]$ into $k$ subsets, and given that $\Pi_k = \{A_1, \ldots, A_k\}$ say, where $\#A_i = n_i$ with $\sum_i n_i = n$, the partition $\Pi_{k-1}$ is derived from $\Pi_k$ by merging $A_i$ and $A_j$ with probability proportional to $\kappa(n_i, n_j)$ for $1 \le i < j \le k$. For the constant kernel $\kappa(x,y) \equiv 1$ this Kingman's [25] coalescent process, which has found extensive applications in genetics. See also [18, 44] for recent studies of other $\mathcal{P}_{[n]}$-valued coalescent Markov chains.

The following proposition gives a formulation in terms of $\mathcal{P}_{[n]}$-valued processes of a result stated without proof by Yao [64, Lemma 1] in a study of average behaviour of set merging algorithms. Following subsections show how various forms of this result have appeared in a variety of other contexts.

**Proposition 6** *Suppose that either*

a) $\Pi_k$ *is the partition of* $[n]$ *generated by the tree components of a random forest obtained by cutting* $k-1$ *bonds at random in a uniform random tree over* $[n]$*, which may be either rooted or unrooted, or*

b) $\Pi_k$ *is the partition of* $[n]$ *into* $k$ *subsets generated by an additive coalescent process* $(\Pi_n, \ldots, \Pi_1)$ *started with* $\Pi_n$ *the partition of* $[n]$ *into singletons.*

*Then for each particular partition* $\{A_1, \ldots, A_k\}$ *of* $[n]$ *into* $k$ *subsets with* $\#A_i = n_i$ *for* $1 \le i \le k$

$$P(\Pi_k = \{A_1, \ldots, A_k\}) = \frac{\prod_{i=1}^k n_i^{n_i - 1}}{n^{n-k} \binom{n-1}{k-1}} \tag{12}$$

**Proof.** By Theorem 5, either a) or b) implies that $\Pi_k$ has the same distribution as the partition generated by a random forest $F_k$ with distribution displayed in (9). But if $\Pi_k$ is so generated by $F_k$, given that $\Pi_k$ equals $\{A_1, \ldots, A_k\}$, the forest $F_k$ is equally likely to be any one of $\prod_{i=1}^k n_i^{n_i - 2}$ possible forests $f_k$. So (12) follows from (9). □

The distribution of the partition of $n$ generated by $\Pi_k$ as above is most simply described by the distribution of the random vector $(N_1^*, \ldots, N_k^*)$ of sizes of components of $\Pi_k$ in *random order*. That is, $N_j^* = N(\sigma_j)$ where $N(1) \geq \ldots \geq N(k)$ are the ranked sizes of components of $\Pi_k$, and $(\sigma_1, \ldots, \sigma_k)$ is a random permutation of $[k]$, assumed independent of $\Pi_k$.

**Proposition 7** *For $\Pi_k$ as in the previous proposition, the distribution of the sizes $N_1^*, \ldots, N_k^*$ of components of $\Pi_k$ in random order is given by*

$$P\left(\bigcap_{i=1}^{k}(N_i^* = n_i)\right) = \frac{(n-k)!}{kn^{n-k-1}} \prod_{i=1}^{k} \frac{n_i^{n_i-1}}{n_i!} \tag{13}$$

*for all $(n_1, \ldots, n_k)$ with $\sum_i n_i = n$.*

**Proof.** It is enough to show this for $\Pi_k$ defined by the tree components of $R_k$, where $R_k$ is the rooted random forest with uniform distribution on $\mathcal{R}_{k,n}$, as in Theorem 4. The distribution of $(N_1^*, \ldots, N_k^*)$ is then unchanged by conditioning the set of roots of trees in $R_k$ to be any particular subset of $k$ elements of $[n]$, say $[k]$, and further unchanged by listing the tree sizes in the deterministic order of their roots. This reduced form of (13), with $N_i^*$ the size of the tree rooted at $i$ in a rooted random forest over $[n]$ with uniform distribution on $\mathcal{R}_{k,n}^0$, is due to Pavlov[39], and can be verified as follows. The number of forests in $\mathcal{R}_{k,n}^0$ in which the tree rooted at $j$ is of size $n_j$ for $1 \leq j \leq k$ and $\sum_{j=1}^{k} n_j = n$ is easily seen to be

$$\frac{(n-k)!}{\prod_{i=1}^{k}(n_i-1)!} \prod_{i=1}^{k} n_i^{n_i-2} \tag{14}$$

Dividing this number by $\#\mathcal{R}_{k,n}^0 = kn^{n-k-1}$ yields (13). □

The observation in the above proof, that uniform random elements of $\mathcal{R}_{k,n}$ and $\mathcal{R}_{k,n}^0$ induce the same distribution of component sizes, was made by Łuczak [29, §5].

## 4.1 Pavlov's representation

As observed by Pavlov[39], the joint distribution of $(N_1^*, \ldots, N_k^*)$ defined by (13) is identical to the conditional distribution of $(N_1, \ldots, N_k)$ given $N_1 + \cdots + N_k = n$ where $N_1, \ldots, N_k$ are independent and identically distributed with the *Borel($\lambda$) distribution*

$$P(N_i = n) = \frac{(n\lambda)^{n-1} e^{-n\lambda}}{n!} \qquad (n = 1, 2, \ldots) \tag{15}$$

for some $0 < \lambda \leq 1$. It well known that such $N_i$ can be constructed by letting $N_i$ be the total progeny of the $i$th of $k$ initial individuals in a Poisson-Galton-Watson branching process in which each individual has $j$ offspring with probability $e^{-\lambda}\lambda^j/j!$, $j = 0, 1, 2, \ldots$. Pavlov [39, 40] applied this representation to obtain a number of results regarding the asymptotic distribution for large $n$ and $k$ of the partition of $n$ induced by such $(N_1^*, \ldots, N_k^*)$. Note that $k$ here is Pavlov's $N$, our $n$ is his $N + n$, our $N_i^*$ is his $\nu_i + 1$, and our $C_j(\Pi_k)$ (introduced in the next subsection) is his $\mu_{j-1}(n-k, k)$. According to Proposition 7, after these translations each of Pavlov's results describes some asymptotic feature of the partition of $n$ generated by $\Pi_k$ as in Proposition 6, in various limiting regimes as both $n$ and $k$ tend to $\infty$. Results of [40] and [2] imply that the random sequence $N_{n,k}(1) \geq N_{n,k}(2) \geq \cdots$ obtained by ranking the sequence of $k$ component sizes of the random partition $\Pi_k$ of $[n]$ is such that the normalized sequence

$$\left( \frac{N_{n,k}(1)}{n}, \frac{N_{n,k}(2)}{n}, \ldots \right)$$

has a non-degenerate limiting distribution parameterized by $\ell > 0$ as $k$ and $n$ tend to $\infty$ with $k \sim \ell\sqrt{n}$. Descriptions of this family of limiting distributions can be read from work of Perman et al. [48, 42, 41, 49]. See [9, 18] for details and further study of the discrete measure valued coalescent process obtained in this limit regime. That the limiting distribution of the normalized sequence exists is an indication that for a random rooted forest to have a giant tree (of size of order $n$), the number $k$ of trees needs to be $O(n^{1/2})$. In contrast, for unrooted forests a giant tree appears much earlier, when the number of trees is about $n/2$ [30].

## 4.2 Randomizing the number of trees

Consider now the distribution of $\Pi_K$, for $(\Pi_n, \ldots, \Pi_1)$ the $\mathcal{P}_{[n]}$-valued additive coalescent process as above, and $K$ a random variable with values in $[n]$, assumed to independent of $(\Pi_n, \ldots, \Pi_1)$. Then from (12) the distribution of $\Pi_K$ on $\mathcal{P}_{[n]}$ is defined by the formula

$$P(\Pi_K = \{A_1, \ldots, A_k\}) = P(K = k) \frac{\prod_{i=1}^k n_i^{n_i - 1}}{n^{n-k} \binom{n-1}{k-1}} \tag{16}$$

for each particular partition $\{A_1, \ldots, A_k\}$ of $[n]$ with $\#A_i = n_i$ for $1 \leq i \leq k \leq n$. For $\Pi \in \mathcal{P}_{[n]}$ let $C_j(\Pi)$ denote the number of components of $\Pi$ of size $j$. Since for each sequence of non-negative integer counts

$$(m_1, \ldots, m_n) \text{ with } \sum_i m_i = k \text{ and } \sum_i i m_i = n$$

the number of partitions $\Pi = \{A_1, \ldots, A_k\} \in \mathcal{P}_{[n]}$ with $C_j(\Pi) = m_j$ for all $1 \le j \le n$ is $n!/(\prod_j j!^{m_j} m_j!)$ and for each of these partitions the probability (16) is the same, it follows from (16) that the probability that the partition of $n$ induced by $\Pi_K$ equals $1^{m_1} 2^{m_2} \cdots n^{m_n}$ is

$$P\left(\bigcap_{j=1}^{n} [C_j(\Pi_K) = m_j]\right) = \frac{P(K=k)\, n!}{\binom{n-1}{k-1} n^{n-k}} \prod_{j=1}^{n} \left(\frac{j^{j-1}}{j!}\right)^{m_j} \frac{1}{m_j!} \tag{17}$$

**Moon's model.** Moon [36] proposed the following model for generating a random forest. For $0 < p < 1$ let $F_{K_p}$ denote the random forest of $K_p$ trees obtained from a uniform random tree $F_1$ over $[n]$ by clipping each of the $n-1$ bonds of $F_1$ independently with the same probability $p$. Then $K_p - 1$, the number of bonds of $F_1$ that are clipped, has binomial$(n-1, p)$ distribution. That is

$$P(K_p = k) = P(K_p - 1 = k - 1) = \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \tag{18}$$

Let $\Pi_{K_p}$ denote the partition of $[n]$ generated by $F_{K_p}$. Formula (17) yields the following expression for the distribution of the partition of $n$ generated by the sizes of tree components of the random forest $F_{K_p}$:

$$P\left(\bigcap_{j=1}^{n} [C_j(\Pi_{K_p}) = m_j]\right) = \frac{n! p^{k-1} (1-p)^{n-k}}{n^{n-k}} \prod_{j=1}^{n} \left(\frac{j^{j-1}}{j!}\right)^{m_j} \frac{1}{m_j!} \tag{19}$$

This formula for a probability distribution over $\mathcal{P}_n$, defined by regarding the right side of (19) as a function of $1^{m_1} 2^{m_2} \cdots n^{m_n} \in \mathcal{P}_n$, was derived from the continuous time $\mathcal{P}_n$-valued additive coalescent process by Lushnikov [32, 31] and Hendriks et al. [22], and discovered in the setting of a two-sex Poisson-Galton-Watson branching process by Sheth [58]. These appearances of this probability distribution on $\mathcal{P}_n$ are described in more detail in following paragraphs.

Let $v$ be any fixed element of $[n]$. Moon [36] found the following formula for the distribution of the random size $J$ of the component of $\Pi_{K_p}$ containing the given vertex $v$:

$$P(J = j) = \frac{p}{1-p} \left(\frac{1-p}{n}\right)^{n-1} \binom{n}{j} j^j \left(\frac{n}{1-p} - j\right)^{n-j-1} \qquad (1 \le j \le n) \tag{20}$$

Sheth [57] derived this probability distribution on $[n]$ from another probabilistic model for clustering whose connection to the present model is explained in [58, 59]. As observed

by Sheth[58], formula (20) follows (19) because given the counts $(C_j(\Pi_{K_p}), 1 \le j \le n)$ the random variable $J$ equals $j$ with probability $jC_j/n$, so $P(J = j) = (j/n)E(C_j(\Pi_{K_p}))$ where $E$ denotes expectation, and formulae for factorial moments of the counts $C_j(\Pi_{K_p})$ can be read from (19) by standard methods.

**Two-sex Poisson-Galton-Watson trees.** Sheth [58, (5)] found the distribution for a partition of $n$ described by (19) by analysis of a stochastic model for gravitational clumping which he reformulated in quite different terms as follows. Consider a Poisson-Galton-Watson branching process (PGW($\lambda$) process) starting with a single male individual, in which each individual has a Poisson($\lambda$) number of offspring for some $0 < \lambda \le 1$, so the process is either critical or sub-critical. Suppose that each individual born in the process is male with probability $p$ and female with probability $q$, independently of the sex of all other individuals. Let $N_{total}$ be the total number of progeny in the branching process, say $N_{total} = N_{male} + N_{female}$. In the random family tree of size $N_{total}$ induced by the branching process, there are $N_{total} - 1$ parent-child bonds. Now cut the tree into subtrees by cutting each bond between a parent and a male child. There are then $N_{male}$ subtrees, each consisting of a single male individual and his female line of descent. Sheth [58] showed that given $N_{total} = n$ the distribution of the partition of $n$ defined by the sizes of the $N_{male}$ distinct subtrees is given by formula which reduces to (19). The connection between Sheth's model and Moon's model is provided by the following observation of Aldous [4] (c.f. Gordon et al. [20], Kolchin [27], Kesten-Pittel [24]). *Given that a* PGW($\lambda$) *branching process starting with one individual has total progeny equal to* $n$*, let* $T$ *be the random tree over* $[n]$ *obtained from the* PGW($\lambda$) *family tree by a random labeling of individuals in the family tree by* $[n]$*. Then* $T$ *has uniform distribution on* $\mathcal{R}_{1,n}$*.* See [47] for a proof and discussion of related results. Combined with Theorem 4 this observation implies that in the two-sex PGW($\lambda$) process in which each child is male with probability $p$, starting with a single male and given that $N_{total} = n$ and $N_{male} = k$, the rooted random forest over $[n]$ defined by subtrees of descent with male roots, for a random labeling of all individuals by $[n]$, has uniform distribution on $\mathcal{R}_{k,n}$. Also, Sheth's result may be reformulated as follows: given that $N_{total} = n$ and $N_{male} = k$, the sequence $(N_1^*, \ldots, N_k^*)$ of sizes of the $k$ subtrees, presented in a random order, has the joint distribution defined by formula (13).

**Random Mappings.** To give one more application of formula (17), let $\phi$ be a random mapping from $[n]$ to $[n]$, with uniform distribution on the set of all $n^n$ such mappings. Let $D$ be the asociated random digraph with the set of $n$ edges $\{i \to \phi(i),\ 1 \le i \le n\}$.

See [28, 2] for background. Let $C$ be the set of vertices in cycles of $D$, let $K = \#C$, and let $R$ be the rooted forest with roots in $C$ obtained by first cutting the edges of $D$ between points in $C$, then reversing all edge directions. So the edges of $R$ are directed away from $C$. It is known [21] that

$$P(K = k) = \frac{k(n-1)!}{n^k(n-k)!} \qquad (1 \le k \le n) \tag{21}$$

and easily seen that given $K = k$ the random forest $R$ has uniform distribution on $\mathcal{R}_{k,n}$. It follows that the random partition of $[n]$ generated by the tree components of $R$ has the same distribution as that of $\Pi_K$ described in (16) and (17) for $K$ with the distribution (21). See Dénes [17] for a similar but more complicated formula for the distribution of the partition of $n$ generated by the components of $D$ rather than $R$, and see [2] regarding the asymptotic joint distribution as $n \to \infty$ of the relative sizes of both kinds of components.

## 4.3 The continuous time additive coalescent.

The random forest $F_{K_p}$ as defined above (18) can be constructed simultaneously for all $0 \le p \le 1$ as follows. Label the edges of a uniform random tree $F_1$ over $[n]$ in an arbitrary way by $i$ with $1 \le i \le n-1$, assign to edge $e_i$ a random variable $U_i$, where the $U_i$ are independent with uniform distribution on $[0,1]$. A refining sequence of forests $(F_1, \ldots, F_n)$ satisfying the conditions of Theorem 5 is then obtained by the cutting edges $e_i$ of $F_1$ in increasing order of the values $U_i$. Let $K_p - 1$ be the number of $i$ with $U_i < p$. Then $K_p$ is independent of $(F_1, \ldots, F_n)$, and $F_{K_p}$ is derived from $F_1$ cutting those $K_p - 1$ edges $e_i$ of $F_1$ with $U_i < p$. Let $\Pi(p)$ be the random partition of $[n]$ generated by the tree components of $F_{K_p}$, and define a process $(\Pi^*(t), 0 \le t < \infty)$ by

$$\Pi^*(t) = \Pi(e^{-t}) \tag{22}$$

That is, $\Pi^*(t)$ is the random partition generated by cutting those edges $e_i$ of $F_1$ such that $W_i > t$, where the $W_i = -\log(U_i)$ are a sequence of $n-1$ independent and identically distributed exponential variables assigned to the edges of $F_1$. Think of $W_i$ as the *birth time* of the edge $e_i$ of the tree $F_1$. Then $\Pi^*(t)$ is the partition generated by tree components in the forest whose edges are those edges of $F_1$ that are born by time $t$.

**Theorem 8** *The process $(\Pi^*(t), 0 \le t < \infty)$ defined by (22) is a continous time Markov chain with state space $\mathcal{P}_{[n]}$ in which at each time $t > 0$, each unordered pair of components of $\Pi^*(t)$ of sizes $x$ and $y$ is merging to form a component of size $x + y$ at rate $(x + y)/n$.*

**Proof.** This follows from the memoryless property of exponential distribution, and the description (ii)' of Theorem 5. □

Theorem 8 implies that formula (19) with $p = e^{-t}$ gives the distribution at time $t$ of the partition of $n$ induced by a coalescent process with collision rate function $\kappa(x, y) := (x + y)/n$ started with the *monodisperse initial condition* $\Pi^*(0) = \{\{1\}, \{2\}, \ldots, \{n\}\}$. It is easily checked that this agrees with the more general result for the distribution at time $t$ of a $\mathcal{P}_n$ valued $\kappa$-coalescent with monodisperse initial condition and collision kernel $\kappa(x, y) = a + b(x + y)$, obtained by Hendriks et al. [22, (19)] by solution of the forwards differential equations. Earlier, Lushnikov [32, 31] obtained an equivalent expression for the additive coalescent in terms of a generating functional. See [59] for applications of Theorem 8 to Sheth's model for gravitational clustering of galaxies. See also [12, 13] for analogous but less explicit results for the multiplicative coalescent.

## 4.4 Random graphs

Let $(G(n, p), 0 \leq p \leq 1)$ denote the usual random graph process with vertex set $[n]$, constructed by assigning each of the $\binom{n}{2}$ possible edges $e$ an independent uniform $[0, 1]$ random variable $U_e$, and letting $G(n, p)$ comprise those edges $e$ with $U_e \leq p$. So for each fixed $p$, the random graph $G(n, p)$ is governed by the model for random graphs commonly denoted $\mathcal{G}(n, p)$. See [10]. As the time parameter $p$ increases from 0 to 1 the random graph process $(G(n, p), 0 \leq p \leq 1)$ develops by addition of edges at random times $0 < P_1 < P_2 < \cdots$ in such a way that each $1 \leq m \leq \binom{n}{2}$, the random graph $G(n, P_m)$ has $m$ edges picked at uniformly at random from the set of $\binom{n}{2}$ possible edges. This model governing $G(n, P_m)$ is commonly denoted $\mathcal{G}(n, m)$. Aldous [3] and Buffet-Pulé [13] studied the random forest process $(F(n, p), 0 \leq p \leq 1)$ derived from $(G(n, p), 0 \leq p \leq 1)$ as follows: whenever $G(n, \cdot)$ adds a new edge, $F(n, \cdot)$ adds the same edge, except if this would create a cycle. As shown in [3], this process $(F(n, p), 0 \leq p \leq 1)$ develops by addition of edges at random times $0 < Q_1 < Q_2 < \cdots < Q_{n-1}$ in such a way that the sequence $(F(n, Q_m), 0 \leq m \leq n - 1$, where $Q_0 = 0$, is a discrete time $\mathcal{F}_n$-valued multiplicative coalescent.

Consider now the dynamic version of Moon's random forest model, say $(F^+(n, u), 0 \leq u \leq 1)$, where $F^+(n, 1)$ is a uniform random tree over $[n]$, and given $F^+(n, 1)$ the forest $F^+(n, u)$ for $0 \leq u \leq 1$ is defined by those bonds of $F^+(n, 1)$ with $U_e \leq u$, where the $U_e$ are independent uniform$(0, 1)$ variables as $e$ ranges over the $n - 1$ edges of $F^+(n, 1)$. Then the process $(F^+(n, u), 0 \leq u \leq 1)$ develops by addition of edges at random times

$0 < U_{(1)} < U_{(2)} < \cdots < U_{(n-1)}$ where the $U_{(m)}$ are the order statistics of the $U_e$. As a consequence of Theorem 4, the sequence $(F^+(n, U_{(m)}), 0 \leq m \leq n-1)$, where $U_{(0)} = 0$, is a discrete time $\mathcal{F}_n$-valued additive coalescent.

Numerous copies of this process $(F^+(n,u), 0 \leq u \leq 1)$ lie embedded in the random graph process $(G(N,p), 0 \leq p \leq 1)$ for any $N \geq n$. Suppose that $C$ is a subset of $[N]$ with $\#C = n$, and that the restriction of $G(N,p)$ to $C$ is a tree. This could be understood by conditioning in various ways. For example, $C$ could be a fixed subset of size $n$, and $G(N,p)$ could be conditioned to make its restriction to $C$ a tree, or to make $C$ a tree component of $G(N,p)$. Or $C$ could be the random component of $G(N,p)$ containing a particular vertex, say vertex 1, and this component could be conditioned to be a tree component of size $n$. Let $G_C(N,q)$ denote the restriction of $G(N,q)$ to $C$, regarded as graph on $[n]$ by relabeling of $C$ via the increasing map from $[n]$ to $C$. Then it follows easily from the basic independence assumptions in the model $\mathcal{G}(N,p)$ that given the existence of the tree component $C$ in $\mathcal{G}(N,p)$ the process $(G_C(N, up), 0 \leq u \leq 1)$ is a copy of the process $(F^+(n,u), 0 \leq u \leq 1)$ defined above. That is to say, *given that a tree component $C$ of size $n$ exists in $G(N,p)$, the development of the sequence of forests that coalesced to form $C$ over the interval $(0,p)$ is governed by an additive coalescent process.* A similar statement holds of course for $G(N,m)$ for arbitrary $m \geq n-1$. Luczak [29, §5] exploited the more obvious time-reversed version of the above statement, in terms of random deletion of edges, to analyze the "race of components" in the evolution of a random graph.

The idea of picking out a clump of a given size $n$ in a coalescent process by suitiable selection or conditioning, and analyzing how this clump of size $n$ was formed, is the idea of a *merger history process* developed in [58, 59]. The point of the above discussion is that even though the overall evolution of the random graph process in its early stages is largely determined by a multiplicative coalescent process [1], the merger history process governing the formation of trees nonetheless involves an additive rather than multiplicative coalescent.

See Pittel [50] and Janson [23] regarding the distribution of the total numbers of tree components of various sizes in a sparse random graph and its dynamic version respectively, and Aldous [1] for recent developments relating the evolution of $(G(n,p), 0 \leq p \leq 1)$, around the critical time $p = 1/n$ of emergence of a giant component, to an infinite version of the multiplicative coalescent.

# 5 Multinomial expansions over trees

The following variation of Moon's derivation of formula (7) yields Cayley's multinomial expansion over trees. Suppose that the partition generated by the tree components of a forest $f_k$ over $[n]$ is $\{A_1, \ldots, A_k\}$ where $\#A_i = n_i$ for $1 \leq i \leq k$ and $\sum_i n_i = n$. To be definite, it will be assumed that the $A_i$ are listed in order of their least elements. Recall that $N(f_k)$ is the number of trees $t \in \mathcal{T}_n$ that contain $f_k$. Moon's formula (7) for $N(f_k)$ is the first of two equalities presented in (23) below. To argue the second equality, observe that each $t \in \mathcal{T}_n$ that contains $f_k$ induces a tree $\tau(f_k, t) \in \mathcal{T}_k$ with a bond from $i$ to $j$ iff $t$ has a bond joining some element of $A_i$ to some element of $A_j$. Call $\tau(f_k, t)$ the *tree induced by* $f_k$ *and* $t$. Given $\tau \in \mathcal{T}_k$ and a forest $f_k$ of $k$ trees over $[n]$, the number of $t \in \mathcal{T}_n$ such that $\tau(f_k, t) = \tau$ is just $\prod_{i=1}^k n_i^{\deg(i,\tau)}$ where $\deg(i, \tau)$ is the *degree of* $i$ in the tree $\tau$, that is the number of bonds of $\tau$ that contain $i$. Thus

$$\left(\prod_{i=1}^k n_i\right) n^{k-2} = N(f_k) = \sum_{\tau \in \mathcal{T}_k} \prod_{i=1}^k n_i^{\deg(i,\tau)} \tag{23}$$

The equality of these two expressions for $N(f_k)$ amounts to

$$(n_1 + \cdots + n_k)^{k-2} = \sum_{\tau \in \mathcal{T}_k} \prod_{i=1}^k n_i^{\deg(i,\tau)-1} \tag{24}$$

This identity, just established for positive integer sequences $(n_i, 1 \leq i \leq k)$ must hold also as an identity of polynomials in $k$ variables $n_i, 1 \leq i \leq k$. This is *Cayley's multinomial expansion over trees*, as formalized by Rényi [53]. Compare coefficients in (24) with those in the usual multinomial expansion to obtain the following equivalent of (24), usually derived either by induction or from the Prüfer coding of trees [38]: for non-negative integers $d_1, \ldots, d_k$ with $\sum_i d_i = 2k - 2$, the number of $\tau \in \mathcal{T}_k$ with $\deg(i, \tau) = d_i$ for all $1 \leq i \leq k$ is the multinomial coefficient

$$\binom{k-2}{d_1 - 1, \ldots, d_k - 1} \tag{25}$$

Recall that if $C_i$ is the number of results $i$ in $n$ independent trials with probability $p_i$ of result $i$ on each trial, where $p_i \geq 0$ and $\sum_{i=1}^k p_i = k$ then the distribution of the vector of counts $(C_1, \ldots, C_k)$ is called *multinomial*$(n; p_1, \ldots, p_k)$. The above enumeration can be restated in probabilistic terms as follows: *Let* $D_1, \ldots, D_k$ *denote the random degrees* $D_i = \deg(i, \tau)$ *for* $\tau$ *picked uniformly at random from* $\mathcal{T}_k$. *Then*

$$(D_1 - 1, \ldots, D_k - 1) \text{ has multinomial } (k-2; k^{-1}, \ldots, k^{-1}) \text{ distribution.} \tag{26}$$

In particular, each $D_i - 1$ has binomial$(k-2, k^{-1})$ distribution, a result due to Clarke [15], and the $D_i - 1$ are asymptotically independent Poisson(1) as $k \to \infty$. There is also the following probabilistic expression of Cayley's multinomial expansion, which reduces to (26) in the special case $n = k$:

**Proposition 9** *For $k \geq 2$ let $F_k$ be the random forest of $k$ trees over $[n]$ obtained by deletion of $k-1$ edges picked at random from the $n-1$ edges of a random tree $F_1$ with uniform distribution on $\mathcal{T}_n$. Given that the partition generated by the tree components of $F_k$ equals $\{A_1, \ldots, A_k\}$, where $\#A_i = n_i$ for $1 \leq i \leq k$ and $\sum_i n_i = n$, let $\tau(F_k, F_1)$ be the random element of $\mathcal{T}_k$ induced by $F_k$ and $F_1$, and let $D_i$ denote the degree of vertex $i$ in the tree $\tau(F_k, F_1)$. Then conditionally given $F_k$,*

$$(D_1 - 1, \ldots, D_k - 1) \text{ has multinomial } \left(k-2; \tfrac{n_1}{n}, \ldots, \tfrac{n_k}{n}\right) \text{ distribution.} \tag{27}$$

**Proof.** For a sequence of positive integers $d_1, \ldots, d_k$ with $\sum_i d_i = 2k-2$, the counting argument leading to (23) shows that for each forest $f_k$ with components of sizes $n_1, \ldots, n_k$, and each particular tree $\tau \in \mathcal{T}_k$ with degree sequence $d_1, \ldots, d_k$,

$$P(\tau(F_k, F_1) = \tau \,|\, F_k = f_k) = \prod_{i=1}^{k} \left(\frac{n_i}{n}\right)^{d_i - 1} \tag{28}$$

Since the number of such trees $\tau$ is given by the multinomial coefficient (25), it follows that

$$P(\cap_{i=1}^{k}(D_i = d_i) \,|\, F_k = f_k) = \binom{k-2}{d_1 - 1, \ldots, d_k - 1} \prod_{i=1}^{k} \left(\frac{n_i}{n}\right)^{d_i - 1} \tag{29}$$

which is equivalent to (27). □

**Corollary 10** *Fix $v \in [n]$. For $k \geq 2$ and $F_k$ and $F_1$ as above, let $A_{k,n}$ be the random subset of $[n]$ defined by the tree component of $F_k$ containing $v$, and let $D_{k,n}$ be the number of edges of $F_1$ which connect $A_{k,n}$ to $[n] - A_{k,n}$. Then*

$$P(\#A_{k,n} = j) = (k-1)\binom{n-k}{j-1}\frac{j^{j-1}(n-j)^{n-j-k}}{n^{n-k}} \qquad (1 \leq j \leq n) \tag{30}$$

*and the conditional distribution of $D_{k,n} - 1$ given $\#A_{k,n} = j$ for $1 \leq j \leq n-k+1$ is binomial with parameters $k-2$ and $j/n$:*

$$P(\#A_{k,n} = j, D_{k,n} - 1 = d) = P(\#A_{k,n} = j)\binom{k-2}{d}\left(\frac{j}{n}\right)^{d}\left(1 - \frac{j}{n}\right)^{k-2-d} \tag{31}$$

**Proof.** By picking a root for $F_1$ uniformly at random, the distribution of $\#A_{k,n}$ can be computed as if $A_{k,n}$ were the component containing $v$ in a uniformly distributed rooted random forest of $k$ trees over $[n]$. An elementary counting argument then gives

$$P(\#A_{k,n} = j) = \binom{n-1}{j-1} j^{j-1} (\#\mathcal{R}_{k-1,n-j})(\#\mathcal{R}_{k,n})^{-1} \tag{32}$$

which yields (30) after substitution of the formula (5) and some cancellation. The binomial distribution of $D_{k,n} - 1$ given $\#A_{k,n}$ can be read from Proposition 9. □

It can be checked using Abel's binomial formula [54]

$$\sum_{i=0}^{N} \binom{N}{i} (x+i)^{i-1} (y+N-i)^{N-i} = x^{-1}(x+y+N)^N \tag{33}$$

that the sum over $1 \le j \le n$ of the probabilities in (30) equals 1. The distribution of $n - k + 1 - \#A_{k,n}$ is that defined by normalization of terms in (33) by their sum for the choice of parameters $N = n - k, x = k - 1, y = 1$. Such a distribution is called *quasi-binomial* [16]. Moon's formula (20) can be recovered by multiplying the probability (30) by the binomial probability (18) and summing over $k$, but the algebra is fairly tedious. See also [43, 45] for study of other distributions related to random trees and Abel's binomial formula.

## 5.1 Expansions over rooted trees

An argument parallel to the above derivation of (23), using Lemma 1 instead of Lemma 3, yields the following variation of Cayley's multinomial expansion as an identity of polynomials in $n_1, \ldots, n_k$ for all $k = 1, 2, \ldots$ :

$$(n_1 + \cdots + n_k)^{k-1} = \sum_{r \in \mathcal{R}_{1,k}} \prod_{i=1}^{k} n_i^{\mathrm{out}(i,r)} \tag{34}$$

where $\mathcal{R}_{1,k}$ is the set of all rooted trees over $[k]$, and $\mathrm{out}(i,r)$, the *out-degree* of $i$ in the rooted tree $r$, is the number of $j$ such that there is an edge $i \to j$ in $r$. That is to say, for non-negative integers $d_1, \ldots, d_k$ with $\sum_i d_i = k - 1$, the number of $r \in \mathcal{R}_{1,k}$ such that $\mathrm{out}(i,t) = d_i$ for all $1 \le i \le k$ is the multinomial coefficient

$$\binom{k-1}{d_1, \ldots, d_k} \tag{35}$$

Moon [38, p. 14] gives a generalization of this formula and indicates a proof by induction. To see (35) more directly, observe that a tree $r \in \mathcal{R}_{1,k}$ with the given out-degrees can be constructed sequentially as follows:

Step 1. Pick which subset of $d_1$ elements of $[k] - \{1\}$ is the set $J_1 = \{j : 1 \to j\}$. The number of possible choices of $J_1$ is $\binom{k-1}{d_1}$.

Step 2. Pick which subset of $d_2$ elements of $[k] - J_1 - \{2\}$ is the set $J_2 = \{j : 2 \to j\}$. No matter what the choice of $J_1$, the number of possible choices is $\binom{k-1-d_1}{d_2}$.

And so on. Note that after the $k-1$ choices which define $r$ there is just one element of $[n]$ that has not been chosen, namely the root of $r$. Multiplying these binomial coefficients gives the multinomial coefficient (35).

The analog of (26) for rooted trees is as follows: *Let $O_1, \ldots, O_k$ denote the random out-degrees $O_i := \mathrm{out}(i, r)$ for $r$ picked uniformly at random from $\mathcal{R}_{1,k}$. Then*

$$(O_1, \ldots, O_k) \text{ has multinomial } (k-1; k^{-1}, \ldots, k^{-1}) \text{ distribution.} \tag{36}$$

In particular, each $O_i$ has binomial$(k-1, 1/k)$ distribution, and the $O_i$ are asymptotically independent Poisson(1) as $k \to \infty$.

The connection between the rooted and unrooted results (36) and (26) can be understood probabilistically as follows. Let the basic probability space be $\mathcal{R}_{1,k} = \mathcal{T}_k \times [k]$ with uniform distribution, and let $\tau$ be the projection of $r = (\tau, x)$ onto $\mathcal{T}_k$. Then the random variables $O_i$ in (36) and $D_i$ in (26) are all defined on the same space $\mathcal{R}_{1,k}$, along with $X$ where $X(\tau, x) = x$ denotes the root of $r = (\tau, x)$. By construction, for each $1 \le i \le k$

$$O_i = (D_i - 1) + 1(X = i) \tag{37}$$

and $X$ has uniform distribution on $[k]$ independent of $(D_1, \ldots, D_k)$. The relation between the results (26) and (36) is now clear, because the term involving $X$ in (37) simply increases the sample size of the multinomial distribution from $k-2$ to $k-1$.

# 6 The additive coalescent semigroup

Theorem 8 described the distribution at time $t$ of a $\mathcal{P}_{[n]}$-valued continuous time Markovian coalescent process with collision kernel $\kappa(x, y) = (x + y)/n$, for initial condition the partition of $[n]$ into singletons. The aim of this section is to obtain a corresponding formula for the distribution at time $t$ of the same coalescent process started at an arbitrary initial partition. That is, to describe the transition semigroup of the additive coalescent. This section provides the combinatorial details of results sketched in Section 3.2 of [18],

which are applied in that paper to establish the existence of additive coalescent processes involving an infinite number of masses.

As in Section 4, the key idea is to first give a more detailed description of a corresponding forest-valued process. The discrete time version of this process is presented in the following generalization of Theorem 4. This result is of some independent interest as it introduces a natural class of non-uniform probability distributions on the set $\mathcal{R}_{k,n}$ of all rooted forests of $k$ trees over $[n]$. See [43, 46] for further study of these distributions.

**Theorem 11** *Let $(p_a, a \in [n])$ be a probability distribution on $[n]$ for some $n \geq 2$. The following two descriptions* (i) *and* (ii) *for the distribution of random sequence $(R_1, R_2, \ldots, R_n)$ of rooted forests over $[n]$ are equivalent and imply for each $1 \leq k \leq n$ that $R_k$ has the probability distribution over $\mathcal{R}_{k,n}$ defined by the formula*

$$P(R_k = r) = \binom{n-1}{k-1}^{-1} \prod_{a=1}^{n} p_a^{\mathrm{out}(a,r)} \qquad (r \in \mathcal{R}_{k,n}); \tag{38}$$

(i) *the tree $R_1$ has the distribution on $\mathcal{R}_{1n}$ defined by* (38) *for $k = 1$, and given $R_1$, for each $1 \leq k \leq n$ the forest $R_k$ is derived by deletion from $R_1$ of $k-1$ edges $e_j, 1 \leq j \leq k-1$, where $(e_j, 1 \leq j \leq n-1)$ is a uniform random permutation of the set of $n-1$ edges of $R_1$;*

(ii) *$R_n$ is the trivial digraph, with $n$ root vertices and no edges, and for $n \geq k \geq 2$, given $R_n, \ldots, R_k$ with $R_k \in \mathcal{R}_{k,n}$, the forest $R_{k-1} \in \mathcal{R}_{k-1,n}$ is derived from $R_k$ by addition of a single directed edge $X_{k-1} \to Y_{k-1}$, where $X_{k-1}$ has distribution $p$, and given also $X_{k-1}$ the vertex $Y_{k-1}$ is picked uniformly at random from the set of $k-1$ roots of the $k-1$ trees in $R_k$ other than the tree containing $X_{k-1}$.*

**Proof.** Fix $n$ and the probability distribution $p$ on $[n]$, and let the sequence $(R_1, R_2, \ldots, R_n)$ be defined by (ii). It will be shown by induction on $k$, starting from $k = n$ and decrementing $k$ by 1 at each step, that formula (38) holds. It is then easily verified that the time-reversed sequence evolves as indicated in (i). For $k = n$ there is a unique forest $r \in \mathcal{R}_{nn}$, so in this case (38) holds by the assumption on $R_n$. Make the inductive hypothesis that (38) holds for $k+1$ instead of $k$, for some $1 \leq k \leq n-1$. That is,

$$P(R_{k+1} = r_{k+1}) = \binom{n-1}{k}^{-1} \Pi(r_{k+1}) \qquad (r_{k+1} \in \mathcal{R}_{k+1,n})$$

where $\Pi(r) := \prod_{a=1}^{n} p_a^{\mathrm{out}(a,r)}$. By construction, for each forest $r_k$ with $k$ tree components which can be obtained by adding a single edge $(x, y)$ to $r_{k+1}$

$$P(R_k = r_k \,|\, R_{k+1} = r_{k+1}) = \frac{p_x}{k}$$

If $r_k$ is any such forest then $\Pi(r_k) = \Pi(r_{k+1})p_x$. From this observation and the inductive hypothesis

$$P(R_{k+1} = r_{k+1}, R_k = r_k) = \frac{1}{k}\binom{n-1}{k}^{-1}\Pi(r_k)$$

The probability $P(R_k = r_k)$ is the sum of these probabilities over all choices of $(r_{k+1}, x, y)$ such that $r_k$ is obtained from $r_{k+1}$ by addition of the edge $(x, y)$. But each $r_k$ with $k$ tree components has $n-k$ edges, so this is a sum of $n-k$ equal terms. Therefore

$$P(R_k = r_k) = \frac{n-k}{k}\binom{n-1}{k}^{-1}\Pi(r_k) = \binom{n-1}{k-1}^{-1}\Pi(r_k)$$

and the induction is complete. □

**Definition 12** For $p$ a probability distribution on $[n]$, call the probability distribution of $R_k$ defined by (38) *the distribution on $\mathcal{R}_{k,n}$ induced by $p$.*

The fact that probabilities in this distribution sum to 1 amounts to the following *multinomial expansion over rooted forests* which is an identity of polynomials in $n$ commuting variables $x_a, a \in [n]$: for each $k \in [n]$

$$\sum_{r \in \mathcal{R}_{k,n}} \prod_{a=1}^{n} x_a^{\mathrm{out}(a,r)} = \binom{n-1}{k-1}\left(\sum_{a=1}^{n} x_a\right)^{n-k} \tag{39}$$

This identity, found independently by Stanley [61], reduces for $k = 1$ to the multinomial expansion over rooted trees in (34). The consequent enumeration of rooted forests by out-degree sequence can be verified by the same method used to check the special case (34). See also [47, 43, 46, 45] for related results.

The coalescent scheme in part (ii) of Theorem 11 provides one simple construction of a random tree $R_1$ with the distribution on $\mathcal{R}_{1,n}$ induced by an arbitrary probability distribution $p$ on $[n]$. Many other constructions of such a tree $R_1$ are possible. For example, the Prüfer coding of rooted trees applied to a sequence of $n-1$ independent random variables with distribution $p$. Or, assuming $p_a > 0$ for all $a$, the following scheme constructs $R_1$ with the distribution on $\mathcal{R}_{1,n}$ induced by $p$ from a sequence of independent random variables $W_0, W_1, \ldots$ with distribution $p$, by application of a general result for irreducible Markov chains [11] [33, §6.1]:

$$R_1 := \{(W_{m-1}, W_m) : W_m \notin \{W_0, \ldots, W_{m-1}\}, m \geq 1\} \tag{40}$$

According to Theorem 11, no matter how a random tree $R_1$ is constructed with the distribution on $\mathcal{R}_{1,n}$ induced by $p$, if $k-1$ edges of $R_1$ are deleted at random, the result is a rooted forest of $k$ trees with the distribution on $\mathcal{R}_{k,n}$ induced by $p$.

Consider now a continuous time version of the process described in Theorem 11. Let $\mathcal{R}_n := \cup_{k=1}^n R_{k,n}$ be the set of all rooted forests over $[n]$.

**Definition 13** Let $p$ be a probability distribution on $[n]$. Call an $\mathcal{R}_n$-valued process $R := (R(t), t \geq 0)$ a *$p$-coalescent forest* if $R$ is a Markov chain such that for each $2 \leq k \leq n$ and each forest $r_k \in \mathcal{R}_{k,n}$, the rate of transitions from $r_k$ to $r$ is zero unless $r \in \mathcal{R}_{k-1,n}$ is derived from $r_k$ by addition of a single edge $x \to y$ for some $y$ which is the root of one of the $k-1$ tree components of $r_k$ that does not contain $x$, and for each such $(x, y)$ the rate of addition of the edge $x \to y$ to $r_k$ is $p_x$.

Note that for any probability distribution $p$ on $[n]$ the set of rooted trees $\mathcal{R}_{1,n}$ is a set of absorbing states for the $p$-coalescent forest, and that for any initial distribution of $R(0)$, in a $p$-coalescent forest the limit $R_1 := R(\infty-)$ exists in $\mathcal{R}_{1,n}$ almost surely. Call this random tree $R_1$ the *terminal random tree* of the $p$-coalescent forest $R$. It follows from the above definition and standard properties of Poisson processes that a $p$-coalescent forest with arbitrary initial state $R(0)$ can be constructed as follows. For each $(x, y) \in [n] \times [n]$ let $N_{x,y}$ be a Poisson process of rate $p_x$, and assume these $N_{x,y}$ are independent. If there is a point of $N_{x,y}$ at time $t$, say that *an edge $(x, y)$ appears at time $t$*. Let $R(t)$ remain the same between times when edges appear. At each time $t$ when an edge $(x, y)$ appears, let $G(t)$ be the directed graph $G(t) := R(t-) \cup \{(x, y)\}$. If $G(t)$ is a rooted forest, let $R(t) = G(t)$, else let $R(t) = R(t-)$.

Suppose now that the initial state $R(0)$ is the trivial forest with $n$ root vertices and no edges. Let $\tau_0 := 0, \tau_n := \infty$. By standard theory of finite state Markov chains, the sequence of times $\tau_1 < \tau_2 < \cdots < \tau_{n-1}$ at which the $p$-coalescent forest changes state is such that the $\tau_i - \tau_{i-1}$ are independent exponential variables with rates $n - i$. Moreover, this sequence is independent of the embedded jumping chain $R_n, \ldots, R_1$ where $R_k$ is the state of $R(t)$ during the interval $[\tau_{n-k}, \tau_{n-k+1})$ when $R(t) \in \mathcal{R}_{k,n}$, and the sequence $(R_k, 1 \leq k \leq n)$ has the distribution described in part (ii) of Theorem 11. For $1 \leq m \leq n-1$ let $(I_m, J_m)$ be the $m$th edge added during the construction of the $p$-coalescent forest $(R(t), t \geq 0)$. Then $((I_m, J_m), 1 \leq m \leq n-1)$ has the same distribution as $((X_{n-m}, Y_{n-m}), 1 \leq m \leq n-1)$ for $((X_j, Y_j), 1 \leq j \leq n-1)$ as in part (ii) of Theorem 11. From the previous discussion, the terminal tree of $(R(t), t \geq 0)$ is

$$R_1 := \{(I_m, J_m), 1 \leq m \leq n-1\} = \lim_{t \to \infty} R(t) \quad \text{a.s.}$$

which has the distribution on $\mathcal{R}_{1,n}$ induced by $p$. According to Theorem 11, conditionally given $R_1$ the sequence $((I_m, J_m), 1 \leq m \leq n-1)$ is a random permutation of the set of $n-1$ edges of $R_1$. For each $(i,j)$ which is an edge of $R_1$, let $\varepsilon_{(i,j)}$ be the random time at which $(i,j)$ is added in the construction of $(R(t), t \geq 0)$; that is $\varepsilon_{(i,j)} = \tau_m$ if $(i,j) = (I_m, J_m)$. Conditionally given $R_1$, the $\varepsilon_{(i,j)}, (i,j) \in R_1$ are a random permutation of $(\tau_1, \ldots, \tau_{n-1})$ where the $\tau_i - \tau_{i-1}$ are independent exponential variables with rates $n-i$, and $(\tau_1, \ldots, \tau_{n-1})$ is the increasing rearrangement of the $(\varepsilon_{(i,j)}, (i,j) \in R_1)$. Conditionally given $R_1$ the joint distribution of the $\varepsilon_{(i,j)}, (i,j) \in R_1$ and the $\tau_1, \ldots, \tau_{n-1}$ is thus identical to the joint distribution of a collection of $n-1$ independent exponential variables with rate 1 and their increasing rearrangement. Thus Theorem 11 implies:

**Corollary 14** *Let $R_1$ be a random tree with the distribution on $\mathcal{R}_{1,n}$ induced by $p$, and independent of $R_1$ let $\varepsilon_j, j \in [n]$ be a collection of independent standard exponential variables. Let*

$$R(t) := \{(i,j) : (i,j) \in R_1, \varepsilon_j \leq t\} \qquad (t \geq 0) \tag{41}$$

*Then $(R(t), t \geq 0)$ is a $p$-coalescent forest with $R(0)$ the trivial forest with no edges, and $R(\infty-) = R_1$ almost surely.*

Since $R_1$ has $n-1$ edges, the number of edges of $R(t)$ is a binomial random variable with parameters $n-1$ and $1-e^{-t}$. Since $R(t) \in \mathcal{R}_{k,n}$ if and only if $R(t)$ has $n-k$ edges,

$$P(R(t) \in \mathcal{R}_{k,n}) = \binom{n-1}{k-1} e^{-(k-1)t}(1-e^{-t})^{n-k} \qquad (1 \leq k \leq n). \tag{42}$$

Hence from (38),

$$P(R(t) = r_k) = e^{-(k-1)t}(1-e^{-t})^{n-k} \prod_{a=1}^{n} p_a^{\mathrm{out}(a,r)} \qquad (r_k \in \mathcal{R}_{k,n}) \tag{43}$$

Consider now the $\mathcal{P}_{[n]}$-valued process $(\Pi(t), t \geq 0)$, where $\Pi(t)$ is the partition of $[n]$ defined by the tree components of $R(t)$. The following lemma is an immediate consequence of the well known criterion in terms of transition rates for a function of a Markov chain to be a Markov chain [56, §IIId]:

**Lemma 15** *If $(R(t), t \geq 0)$ is a $p$-coalescent forest, then $(\Pi(t), t \geq 0)$ is a $\mathcal{P}_{[n]}$-valued additive $p$-coalescent, as per the next definition.*

**Definition 16** Call a $\mathcal{P}_{[n]}$-valued process $(\Pi(t), t \geq 0)$ a $\mathcal{P}_{[n]}$-*valued additive p-coalescent* or $(\mathcal{P}_{[n]}, +, p)$ *coalescent* if $(\Pi(t), t \geq 0)$ is a Markov chain with the following transition rates: for a partition $\pi = \{S_1, \ldots, S_k\} \in \mathcal{P}_{[n]}$ with $k \geq 1$, the only possible transitions out of state $\pi$ are to $\pi_{ij}$ for some $1 \leq i < j \leq k$, where $\pi_{ij}$ is obtained from $\pi$ by merging $S_i$ and $S_j$ and leaving the other components of $\pi$ unchanged, and the rate of transitions from $\pi$ to $\pi_{ij}$ is $p_{S_i} + p_{S_j}$, where $p_S = \sum_{a \in S} p_a$.

Intuitively, think of $p$ as a distribution of mass over $S$. The $(\mathcal{P}_{[n]}, +, p)$-coalescent develops by merging each pair of components at a rate equal to the sum of the masses of the two components. For $p$ the uniform distribution on $\mathcal{P}_{[n]}$, the $(\mathcal{P}_{[n]}, +, p)$-coalescent is identical to the $\mathcal{P}_{[n]}$-valued additive coalescent process considered in Theorem 8. The above development now yields the following description of the semigroup of the $(\mathcal{P}_{[n]}, +, p)$-coalescent for an arbitrary probability measure $p$ on $[n]$:

**Theorem 17** [18] *Let $(\Pi(t), t \geq 0)$ be a $(\mathcal{P}_{[n]}, +, p)$-coalescent, with initial state $\Pi(0)$ the partition of $[n]$ into singletons. Then*

$$P(\Pi(t) = \{S_1, \ldots, S_k\}) = e^{-(k-1)t}(1 - e^{-t})^{n-k} \prod_{i=1}^{k} p_{S_i}^{|S_i|-1} \tag{44}$$

*where $p_S = \sum_{s \in S} p_s$ and $|S|$ is the number of elements of $S$. If instead the initial partition is $\Pi(0) = \pi \in \mathcal{P}_{[n]}$, the same formula applies to each partition $\{S_1, \ldots, S_k\}$ of $[n]$ such that each $S_i$ is a union of some number $n_i$ of components of $\pi$, with $|S_i|$ replaced by $n_i$.*

**Proof.** For $\Pi(0)$ the partition of $[n]$ into singletons, by application of Lemma 15 the probability of the event that $\Pi(t) = \{S_1, \ldots, S_k\}$ is obtained by summing the expression (43) over all forests $r$ whose tree components are $S_1, \ldots, S_k$. Write the product over $S$ in (42) as the product over $1 \leq i \leq k$ of products over $S_i$. The sum of products is then a product of sums, where the $i$th sum is a sum over all trees labelled by $S_i$. Each of these sums can be evaluated by the multinomial expansion over rooted trees (34), and the result is (44). If instead $\Pi(0) = \pi = \{A_1, \ldots, A_q\}$ say, then the only partitions $\{S_1, \ldots, S_k\}$ which are possible states of $\Pi(t)$ are coarsenings of the initial partition. Every such coarsening is identified in an obvious way by a partition of the set $[q]$. With this identification the $(\mathcal{P}_{[n]}, +, p)$ coalescent with initial partition $\{A_1, \ldots, A_q\} \in \mathcal{P}_{[n]}$ is identified with the $(\mathcal{P}_{[q]}, +, p')$ coalescent with initial state the partition of $[q]$ into singletons, and $p'_i = p_{A_i}, i \in [q]$, and the conclusion follows.

# Acknowledgments

Thanks to Ravi Sheth for showing me his work [58] which was the inspiration for this paper, to Boris Pittel for drawing my attention to the paper of Yao [64], and to David Aldous and Steven Evans for many stimulating conversations.

# References

[1] D. Aldous. Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Probab.*, 25:812–854, 1997.

[2] D. Aldous and J. Pitman. Brownian bridge asymptotics for random mappings. *Random Structures and Algorithms*, 5:487–512, 1994.

[3] D.J. Aldous. A random tree model associated with random graphs. *Random Structures Algorithms*, 1:383–402, 1990.

[4] D.J. Aldous. The continuum random tree I. *Ann. Probab.*, 19:1–28, 1991.

[5] 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.

[6] D.J. Aldous. The continuum random tree III. *Ann. Probab.*, 21:248–289, 1993.

[7] D.J. Aldous. Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Probab.*, 25:812–854, 1997.

[8] D.J. Aldous. Deterministic and stochastic models for coalescence: a review of the mean-field theory for probabilists. To appear in *Bernoulli*. Available via homepage http://www.stat.berkeley.edu/users/aldous, 1997.

[9] D.J. Aldous and J. Pitman. The standard additive coalescent. Technical Report 489, Dept. Statistics, U.C. Berkeley, 1997. To appear in *Ann. Probab.*. Available via http://www.stat.berkeley.edu/users/pitman.

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

[11] A. Broder. Generating random spanning trees. In *Proc. 30'th IEEE Symp. Found. Comp. Sci.*, pages 442–447, 1989.

[12] E. Buffet and J.V. Pulé. On Lushnikov's model of gelation. *J. Statist. Phys.*, 58:1041–1058, 1990.

[13] E. Buffet and J.V. Pulé. Polymers and random graphs. *J. Statist. Phys.*, 64:87–110, 1991.

[14] 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).

[15] L.E. Clarke. On Cayley's formula for counting trees. *J. London Math. Soc.*, 33:471–474, 1958.

[16] P.C. Consul. A simple urn model dependent upon a predetermined strategy. *Sankhya Ser. B*, 36:391–399, 1974.

[17] J. Dénes. On a generalization of permutations: some properties of transformations. In *Permutations: Actes du colloque sur les permutations; Paris, Univ. René-Descartes, 10-13 juillet 1972*, pages 117–120. Gauthier-Villars, Paris-Bruxelles-Montréal, 1972.

[18] 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*.

[19] S. Glicksman. On the representation and enumeration of trees. *Proc. Camb. Phil. Soc.*, 59:509–517, 1963.

[20] M. Gordon, T.G. Parker, and W.B. Temple. On the number of distinct orderings of a vertex-labeled graph when rooted on different vertices. *J. Comb. Theory*, 11:142–156, 1971.

[21] B. Harris. Probability distributions related to random mappings. *Ann. Math. Statist.*, 31:1045–1062, 1960.

[22] E.M. Hendriks, J.L. Spouge, M. Eibl, and M. Shreckenberg. Exact solutions for random coagulation processes. *Z. Phys. B - Condensed Matter*, 58:219–227, 1985.

[23] S. Janson. The minimal spanning tree in a complete graph and a functional limit theorem for trees in a random graph. *Random Structures and Algorithms*, 7:337–355, 1995.

[24] H. Kesten and B. Pittel. A local limit theorem for the number of nodes, the height, and the number of leaves in a critical branching process tree. *Random Structures and Algorithms*, 8:243–299, 1996.

[25] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

[26] D. E. Knuth and A. Schonhage. The expected linearity of a simple equivalence algorithm. *Theor. Computer Sci.*, 6:281–315, 1978.

[27] V.F. Kolchin. Branching processes, random trees, and a generalized scheme of arrangements of particles. *Mathematical Notes of the Acad. Sci. USSR*, 21:386–394, 1977.

[28] V.F. Kolchin. *Random Mappings.* Optimization Software, New York, 1986. (Translation of Russian original).

[29] T. Luczak. Component behavior near the critical point of the random graph process. *Random Structures and Algorithms*, 1:287–310, 1990.

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

[31] A.A. Lushnikov. Certain new aspects of the coagulation theory. *Izv. Atmos. Ocean Phys.*, 14:738–743, 1978.

[32] A.A. Lushnikov. Coagulation in finite systems. *J. Colloid and Interface Science*, 65:276–285, 1978.

[33] R. Lyons and Y. Peres. Probability on trees and networks. Book in preparation, available at http://www.ma.huji.ac.il/ lyons/prbtree.html, 1996.

[34] A.H. Marcus. Stochastic coalescence. *Technometrics*, 10:133 – 143, 1968.

[35] J. W. Moon. Enumerating labelled trees. In F. Harary, editor, *Graph Theory and Theoretical Physics*, pages 261–272. Academic Press, 1967.

[36] J. W. Moon. A problem on random trees. *J. Comb. Theory B*, 10:201–205, 1970.

[37] J.W. Moon. Various proofs of Cayley's formula for counting trees. In F. Harary, editor, *A Seminar on Graph Theory*, pages 70–78. Holt, Rineharte and Winston, New York, 1967.

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

[39] Yu. L. Pavlov. Limit theorems for the number of trees of a given size in a random forest. *Math. USSR Subornik*, 32:335–345, 1977.

[40] Yu. L. Pavlov. The asymptotic distribution of maximum tree size in a random forest. *Theory of Probability and its Applications*, 22:509–520, 1977.

[41] M. Perman. Order statistics for jumps of normalized subordinators. *Stoch. Proc. Appl.*, 46:267–281, 1993.

[42] M. Perman, J. Pitman, and M. Yor. Size-biased sampling of Poisson point processes and excursions. *Probab. Th. Rel. Fields*, 92:21–39, 1992.

[43] 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.

[44] J. Pitman. Coalescents with multiple collisions. Technical Report 495, Dept. Statistics, U.C. Berkeley, 1997.
Available via http://www.stat.berkeley.edu/users/pitman.

[45] J. Pitman. The asymptotic behavior of the Hurwitz binomial distribution. Technical Report 500, Dept. Statistics, U.C. Berkeley, 1997. Available via http://www.stat.berkeley.edu/users/pitman.

[46] J. Pitman. The multinomial distribution on rooted labeled forests. Technical Report 499, Dept. Statistics, U.C. Berkeley, 1997. Available via http://www.stat.berkeley.edu/users/pitman.

[47] 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.

[48] J. Pitman and M. Yor. Arcsine laws and interval partitions derived from a stable subordinator. *Proc. London Math. Soc. (3)*, 65:326–356, 1992.

[49] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, 25:855–900, 1997.

[50] B. Pittel. On tree census and the giant component in sparse random graphs. *Random Structures and Algorithms*, 1:311–342, 1990.

[51] H. Prüfer. Neuer Beweis eines Satzes über Permutationen. *Archiv für Mathematik und Physik*, 27:142–144, 1918.

[52] A. Rényi. Some remarks on the theory of trees. *MTA Mat. Kut. Inst. Kozl.*, 4:73–85, 1959.

[53] 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.

[54] J. Riordan. *Combinatorial Identities*. Wiley, New York, 1968.

[55] J. Riordan. Forests of labeled trees. *J. Comb. Theory*, 5:90–103, 1968.

[56] M. Rosenblatt. *Random Processes*. Springer-Verlag, New York, 1974.

[57] R.K. Sheth. Merging and hierarchical clustering from an initially Poisson distribution. *Mon. Not. R. Astron. Soc.*, 276:796–824, 1995.

[58] R.K. Sheth. Galton-Watson branching processes and the growth of gravitational clustering. *Mon. Not. R. Astron. Soc.*, 281:1277–1289, 1996.

[59] R.K. Sheth and J. Pitman. Coagulation and branching process models of gravitational clustering. *Mon. Not. R. Astron. Soc.*, 289:66–80, 1997.

[60] P.W. Shor. A new proof of Cayley's formula for counting labelled trees. *J. Combinatorial Theory A.*, 71:154–158, 1995.

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

[62] R. P. Stanley. *Enumerative Combinatorics, Vol I.* Wadsworth & Brooks/Cole, Monterey, California, 1986.

[63] L. Takács. Counting forests. *Discrete Mathematics*, 84:323–326, 1990.

[64] A. Yao. On the average behavior of set merging algorithms. In *Proc. 8th ACM Symp. Theory of Computing*, pages 192–195, 1976.

[65] R. M. Ziff. Kinetics of polymerization. *J. Stat. Physics*, 23:241–263, 1980.