A family of random trees with random edge lengths*

by David Aldous and Jim Pitman

Technical Report No. 526

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

October 14, 1998

Abstract

We introduce a family of probability distributions on the space of trees with I labeled vertices and possibly extra unlabeled vertices of degree 3, whose edges have positive real lengths. Formulas for distributions of quantities such as degree sequence, shape, and total length are derived. An interpretation is given in terms of sampling from the inhomogeneous continuum random tree of Aldous and Pitman (1998).

Key words and phrases. Continuum tree, enumeration, random tree, spanning tree, weighted tree, Cayley's multinomial expansion.
AMS 1991 subject classification. 05C05, 60C05

*Research supported in part by N.S.F. Grants DMS 96-22859 and 97-03961

1 Introduction

By a *discrete tree* we mean a finite tree in the usual sense of graph theory: n vertices connected by $n-1$ undirected edges. By a *tree with edge lengths* we mean a discrete tree in which each edge is assigned a strictly positive real number, which we interpret as the *length* of the edge. Such trees are often called *weighted trees*, but we wish to emphasize our interpretation of the weights as edge lengths. Study of the properties of random discrete trees, which for uniform models of randomness amounts to enumerations of various sets of trees, is a classical topic [5, 11, 13, 14]. Probability models for random trees with edge lengths arise in two specific settings.

(a) Minimum spanning trees and Steiner trees on random points in d-dimensional space; here the edge-lengths are ordinary Euclidean lengths [9, 10, 22].

(b) Genealogical trees representing ancestry of individuals in a species, or phylogenetic trees representing evolutionary relationships between species; here the edge-lengths represent times between divergence of lineages [12, 23].

Another model for such trees is to start from a model of random discrete trees and assign i.i.d. edge lengths [8]. The purpose of this paper is to describe a new model of random trees with edge lengths. This model arises from the study of the asymptotic sizes and shapes of spanning subtrees in a model for random discrete trees studied in [16, 17, 20]. The model is parametrized by a vector (c_i) of vertex weights. A rough interpretation of c_i is the relative propensity of vertex i to have incident edges. Varying these parameters will vary the typical shape of realized trees.

The model is defined in Theorem 1. Section 3.2 shows how it arises as a limit of a natural model for random discrete trees. Our emphasis is on obtaining explicit distributional formulas for quantities associated with the random tree. But we also note (section 7) an interpretation of our model in terms of the *inhomogeneous continuum random tree* (ICRT) introduced in [2] as the key to analysis of a certain continuous-space Markov process. This interpretation provides additional motivation for studying the model, but is not essential for understanding the results of this paper.

2 Overview of results

To state our results we first need some notation for spaces of trees. For a finite set F let T_F be the set of discrete trees whose vertex set consists of labeled vertices F, called *hubs* and perhaps extra unlabeled vertices of degree exactly 3, called *junctions*. Given $\mathbf{t} \in T_F$ for some finite set F, write $\mathcal{E}(\mathbf{t})$ for its set of edges. Write $D_i\mathbf{t}$ for the degree of i in $\mathbf{t}$. If $D_i\mathbf{t} = 1$ call i a *leaf* of $\mathbf{t}$. Assigning a vector $\boldsymbol{l} := (l_e, e \in \mathcal{E}(\mathbf{t}))$ of strictly positive

real lengths to the edges of a tree $\mathbf{t}$ in T_F gives a *tree with edge-lengths*, say $\mathbf{s}$, with $\text{shape}(\mathbf{s}) = \mathbf{t}$ and $\text{lengths}(\mathbf{s}) = \boldsymbol{l}$. Write $\mathbf{T}_F$ for the set of such trees with edge-lengths. Let $[n] := \{1, 2, \ldots, n\}$. Figure 1 shows an element $\mathbf{s}$ of $\mathbf{T}_{[8]}$ with 8 hubs, 6 leaves and 2 junctions. In such a diagram the location of vertices in the plane is arbitrary subject to the shape of the tree and its edge lengths.

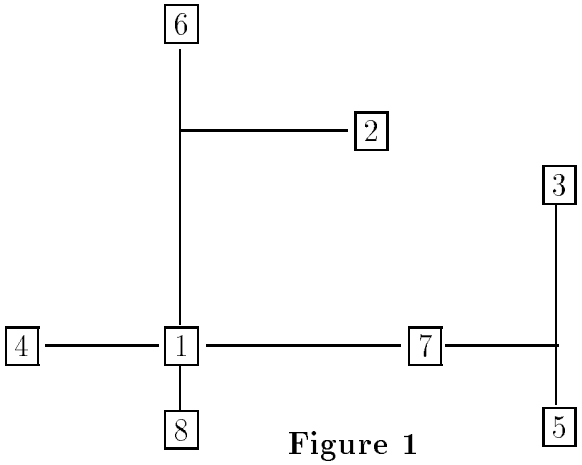

Figure 1

The subject of this paper is the distribution on $\mathbf{T}_{[I]}$ defined in Theorem 1. For $I = 2, 3, \ldots$ let

$$\mathcal{C} := \{\mathbf{c} := (c_1, c_2, \ldots, c_I) : \ I \geq 2, c_i \geq 0 \text{ for each } 1 \leq i \leq I\}.$$

The distribution is parametrized by $\mathbf{c} \in \mathcal{C}$.

Theorem 1 *For each $\mathbf{c} = (c_1, \ldots, c_I) \in \mathcal{C}$, the following formula defines a probability distribution for a random $\mathbf{T}_{[I]}$-valued tree $\mathcal{S}_\mathbf{c}$:*

$$\begin{aligned} &P(\text{shape}(\mathcal{S}_\mathbf{c}) = \mathbf{t}, \text{lengths}(\mathcal{S}_\mathbf{c}) \in [\boldsymbol{l}, \boldsymbol{l} + d\boldsymbol{l}]) \\ &\quad = \left(\prod_{i=1}^{I} c_i^{D_i \mathbf{t} - 1} \right) \ (s + c) \exp(-\tfrac{1}{2} s^2 - sc) \ d\boldsymbol{l}, \quad \mathbf{t} \in T_{[I]}, \ \boldsymbol{l} \in (0, \infty)^{\mathcal{E}(\mathbf{t})} \end{aligned} \tag{1}$$

where $\boldsymbol{l} := (l_e, e \in \mathcal{E}(\mathbf{t}))$, $s := \sum_{e \in \mathcal{E}(\mathbf{t})} l_e$ and $c := \sum_{i=1}^{I} c_i$.

If a constant of normalization, say $Z(c_1, \ldots, c_I)$, were introduced on the right side of formula (1), then the conclusion of Theorem 1 would be obvious. So the content of Theorem 1 is that $Z(c_1, \ldots, c_I) \equiv 1$. For instance, if $I = 2$ the tree $\mathcal{S}_\mathbf{c}$ has a single edge, whose length has probability density function $s \to (s + c) \exp(-\frac{1}{2} s^2 - sc)$ where

$c := c_1 + c_2$. In this case a simple integration verifies this is a probability density. In section 3.2 we give a discrete approximation argument which shows how the formula (1) arises. The proof of Theorem 1 is completed in section 4.4.

Call (1) the *basic formula*. Note that we allow $c_i = 0$, in which case we interpret $0^0 = 1$. Then (1) implies

$$\text{if } c_i = 0 \text{ then vertex } i \text{ is a leaf in } \mathcal{S}_{\mathbf{c}}. \tag{2}$$

We choose the symbol $\mathcal{S}$ for this type of random tree with edge lengths partly by analogy with *Steiner* trees (which also have the feature of extra degree-3 vertices), and partly because we will later interpret $\mathcal{S}$ as a *spanning subtree* within an ICRT.

Our emphasis is on obtaining explicit distributional formulas for quantities associated with the random tree $\mathcal{S}_{\mathbf{c}}$. For $\mathbf{t} \in T_{[I]}$ we shall consider the total edge length $L(\mathbf{t}) := \sum_{e \in \mathcal{E}(\mathbf{t})} l_e$ and the total excess degree $D(\mathbf{t}) := \sum_{i=1}^{I} (D_i \mathbf{t} - 1)$. We shall give explicit formulas for the distributions of

- $L(\mathcal{S}_{\mathbf{c}})$: Corollary 9(i)
- $D(\mathcal{S}_{\mathbf{c}})$: Proposition 10(ii)
- shape$(\mathcal{S}_{\mathbf{c}})$: Proposition 7

and associated joint distributions. The derivations use an enumeration of trees in $T_{[I]}$ by degree sequence, Proposition 8.

Many natural questions involve the subtree of $\mathcal{S}_{\mathbf{c}}$ spanned by some subset V of two or more elements of $[I]$. Denote this subtree of $\mathcal{S}_{\mathbf{c}}$ by $\mathcal{S}_{\mathbf{c}}^V$. For example, the distance between i and j in $\mathcal{S}_{\mathbf{c}}$ is the length of $\mathcal{S}_{\mathbf{c}}^{\{i,j\}}$. Whether or not k is on the path from i to j is a question involving the shape of $\mathcal{S}_{\mathbf{c}}^{\{i,j,k\}}$, and so on. Let hubs$(\mathcal{S}_{\mathbf{c}}^V)$ be the set of labeled vertices of $\mathcal{S}_{\mathbf{c}}^V$. So hubs$(\mathcal{S}_{\mathbf{c}}^V)$ is a random set with $V \subseteq \text{hubs}(\mathcal{S}_{\mathbf{c}}^V) \subseteq [I]$. Given that hubs$(\mathcal{S}_{\mathbf{c}}^V) = H$ regard $\mathcal{S}_{\mathbf{c}}^V$ as a random element of $\mathbf{T}_H$.

The distribution of $\mathcal{S}_{\mathbf{c}}^V$ is determined by the following theorem.

Theorem 2 *For $\mathbf{c} \in \mathcal{C}$ let $\mathcal{S}_{\mathbf{c}}$ be a $\mathbf{T}_{[I]}$-valued random tree with the distribution defined by the basic formula (1). Let $\mathcal{S}_{\mathbf{c}}^V$ be the subtree of $\mathcal{S}_{\mathbf{c}}$ spanned by some subset V of two or more elements of $[I]$. Then for each H with $V \subseteq H \subseteq [I]$ and each $\mathbf{t} \in T_H$ such that $\mathbf{t}$ is spanned by V*

$$\begin{aligned} &P\left(\text{hubs}(\mathcal{S}_{\mathbf{c}}^V) = H, \text{shape}(\mathcal{S}_{\mathbf{c}}^V) = \mathbf{t}, \text{lengths}(\mathcal{S}_{\mathbf{c}}^V) \in [\boldsymbol{l}, \boldsymbol{l} + d\boldsymbol{l}]\right) \\ &\quad = \left(\prod_{h \in H} c_h^{D_h \mathbf{t} - 1}\right) \; (s + c_H) \exp(-\tfrac{1}{2}s^2 - sc) \; d\boldsymbol{l}, \; \boldsymbol{l} \in (0, \infty)^{\mathcal{E}(\mathbf{t})} \end{aligned} \tag{3}$$

where $\boldsymbol{l} := (l_e, e \in \mathcal{E}(\mathbf{t}))$, $s := \sum_{e \in \mathcal{E}(\mathbf{t})} l_e$, $c_H := \sum_{h \in H} c_h$, and $c := c_{[I]}$.

Call (3) the *master formula*. Note that for $V = H$ the master formula reduces to the basic formula. Though the master formula is in principle determined by the basic formula via appropriate summations and integrations, these sums and integrals are not easy to evaluate except in special cases. Rather, the master formula is derived in section 3.2 by a discrete approximation argument which parallels the derivation of the basic formula.

To illustrate a consequence of the master formula, consider for $j, k \in [I]$ the distance $L_{jk}(\mathcal{S}_{\mathbf{c}})$ between vertices j and k in $\mathcal{S}_{\mathbf{c}}$. In section 6 we obtain the following remarkable formula:

Corollary 3 *For distinct $j, k \in [I]$ and $s > 0$*

$$P(L_{jk}(\mathcal{S}_{\mathbf{c}}) > s) = e^{-\frac{1}{2}s^2 - sc} \prod_{i \in [I] \setminus \{j,k\}} (1 + c_i s). \tag{4}$$

That is to say, $L_{jk}(\mathcal{S}_{\mathbf{c}})$ has the same distribution as the minimum of $I + 1$ independent random variables $W_i, 0 \leq i \leq I$ where W_0 has the Rayleigh distribution $P(W_0 > s) = e^{-\frac{1}{2}s^2}$, while W_i has the exponential(c_i) distribution $P(W_i > s) = e^{-c_i s}$ for $i \in \{j, k\}$ and the gamma$(2, c_i)$ distribution $P(W_i > s) = e^{-c_i s}(1 + c_i s)$ for $i \notin \{j, k\}$. Only in the simplest case when $c_j = c_k = 0$ are we able to give a direct probabilistic derivation of this result. This derivation, given in section 7.3, is based on a construction of $\mathcal{S}_{\mathbf{c}}$ related to the interpretation of this random tree as a subtree of an ICRT.

3 Discrete trees and the convergence argument

3.1 Inhomogeneous random discrete trees

We quote two results about discrete trees. For a finite set A write $\#A$ for the number of elements of A, and write U_A for the set of all $(\#A)^{\#A-2}$ discrete trees with vertex-set A.

Lemma 4 [17] *Associated with each probability distribution $p = (p_a)$ on a finite set A is a probability distribution on U_A:*

$$P(\mathcal{U} = \mathbf{u}) = \prod_{a \in A} p_a^{D_a \mathbf{u} - 1} \tag{5}$$

where $D_a\mathbf{u}$ is the degree of a in $\mathbf{u}$.

Call $\mathcal{U}$ a p-tree. As discussed in Pitman [17], the fact that (5) defines a probability distribution without any extra normalization constant amounts to Cayley's multinomial expansion over trees. See [16, 17, 19, 20] for various applications of p-trees and associated random forests.

Lemma 5 [17] *Let $\mathcal{U}$ be a p-tree labeled by a finite set A. Let $\mathcal{U}_F$ denote the subtree of $\mathcal{U}$ spanning a subset F of A with $\#F \geq 2$. Then for every tree $\mathbf{u}$ labeled by a finite subset $V(\mathbf{u})$ of A, such that $\mathbf{u}$ is spanned by F,*

$$P(\mathcal{U}_F = \mathbf{u}) = \left(\prod_{v \in V(\mathbf{u})} p_v^{D_v \mathbf{u} - 1} \right) \left(\sum_{v \in V(\mathbf{u})} p_v \right). \tag{6}$$

3.2 The convergence argument

We now show how the basic formula and the master formula follow from Lemmas 4 and 5 by discrete approximation.

Fix $\mathbf{c} = (c_1, \ldots, c_I) \in \mathcal{C}$ and recall $\sum_i c_i = c$. For sufficiently large n, define a probability distribution $p^{[n]}$ on $[n]$ by

$$\begin{aligned} p_i^{[n]} &= \max(c_i/n^{1/2}, 1/n^{3/4}), \ 1 \leq i \leq I \\ &= q_n, \ I+1 \leq i \leq n \end{aligned} \tag{7}$$

where $q_n = (1 - \sum_1^I p_i^{[n]})/(n-I)$. Let $\mathcal{U}_n$ be the associated $U_{[n]}$-valued random $p^{[n]}$-tree. Write $\mathbf{s}(\mathcal{U}_n)$ for the subtree of $\mathcal{U}_n$ spanned by $[I]$. We regard $\mathbf{s}(\mathcal{U}_n)$ as taking values in $\bar{\mathbf{T}}_{[I]}$, where $\bar{\mathbf{T}}_{[I]}$ is defined like $\mathbf{T}_{[I]}$ but with the condition that unlabeled vertices have degree 3 replaced by the condition that unlabeled vertices have degree 3 or more. Thus for each $\mathbf{u} \in U_{[n]}$ the tree $\mathbf{s}(\mathbf{u}) \in \bar{\mathbf{T}}_{[I]}$ is defined as follows. First, let $\mathbf{s}'(\mathbf{u})$ be the subtree of $\mathbf{u}$ spanned by $[I]$, regarded as an element of U_F where F with $[I] \subseteq F \subseteq [n]$ is the set of all vertices of $\mathbf{u}$ which lie on the path in $\mathbf{u}$ joining i and j for some $i, j \in [I]$. Let J be the set of all $j \in [n] \setminus [I]$ such that j is a vertex of degree 3 or more in $\mathbf{s}'(\mathbf{u})$. Let $\mathbf{s}''(\mathbf{u})$ be the tree in $\bar{\mathbf{T}}_{[I] \cup J}$, with all labeled vertices, with an edge joining i to j of length m iff there is a path of m edges of $\mathbf{u}$ from i to j via $m-1$ vertices of degree 2 in $\mathbf{s}'(\mathbf{u})$. Finally, let $\mathbf{s}(\mathbf{u}) \in \bar{\mathbf{T}}_{[I]}$ be $\mathbf{s}''(\mathbf{u})$ with all vertices in J delabeled.

Proposition 6 *Fix $\mathbf{t} \in T_{[I]}$ and $\boldsymbol{l}^* = (l_e^*, e \in \mathcal{E}(\mathbf{t}))$. Let $s := \sum_{e \in \mathcal{E}(\mathbf{t})} l_e$. As $n \to \infty$ and $\boldsymbol{l}$ ranges over vectors of positive integers with $n^{-1/2}\boldsymbol{l} \to \boldsymbol{l}^*$,*

$$P(\mathrm{shape}(\mathbf{s}(\mathcal{U}_n)) = \mathbf{t}, \mathrm{lengths}(\mathbf{s}(\mathcal{U}_n)) = \boldsymbol{l}) \tag{8}$$

$$= \left(\left(\prod_{i=1}^{I} c_i^{D_i \mathbf{t} - 1} \right) (s+c) e^{-\frac{1}{2}s^2 - sc} + o(1) \right) \ n^{-\#\mathcal{E}(\mathbf{t})/2}. \tag{9}$$

Proof. For a given $\mathbf{t} \in T_{[I]}$ and lengths $\boldsymbol{l} = (l_e, e \in \mathcal{E}(\mathbf{t}))$ each unrooted tree $\mathbf{u}$ labeled by $[n]$ such that $\mathrm{shape}(\mathbf{s}(\mathbf{u})) = \mathbf{t}$ and $\mathrm{lengths}(\mathbf{s}(\mathbf{u})) = \boldsymbol{l}$ has the same number of vertices

in $[n] \setminus [I]$, say v which is given by $v = \sum_e l_e + 1 - I$, and the same number of junction vertices in $[n] - [I]$, say j. Since the number of such trees $\mathbf{u}$ is

$$(n-I)_v := (n-I)(n-I-1)\cdots(n-I-v+1)$$

by application of Lemma 5, the probability in (8) equals

$$(n-I)_v \left[\prod_{i=1}^{I} \left(p_i^{[n]}\right)^{D_i\mathbf{t}-1}\right] q_n^{v+j} \left(\sum_{i=1}^{I} p_i^{[n]} + vq_n\right). \tag{10}$$

Let $d_I := \sum_{i=1}^{I}(D_i\mathbf{t}-1)$. Note that

$$(p_i^{[n]})^{D_i\mathbf{t}-1} = (c_i^{D_i\mathbf{t}-1} + o(1))n^{-(D_i\mathbf{t}-1)/2}.$$

In the limit regime with $v \sim s\sqrt{n}$ the expression (10) is asymptotically

$$\left(\left[\prod_{i=1}^{I} c_i^{D_i\mathbf{t}-1}\right] \frac{(n-I)_v}{(n-I)^v}(1-c/\sqrt{n})^{v+j}(c+v/\sqrt{n}) + o(1)\right) n^{-(d_I+2j+1)/2}.$$

By (16,17) we have $d_I + 2j + 1 = \#\mathcal{E}(\mathbf{t})$. Since $v + j \sim s\sqrt{n}$ this expression is asymptotically equivalent to that displayed in (9). □

Proof of Theorems 1 and 2. The coefficient of $n^{-\#\mathcal{E}(\mathbf{t})/2}$ on the right side of (9) is the density (1). Since the left side of (9) is a probability measure, it easily follows that (1) is the density of a measure with total mass $\mu \leq 1$, and that the property $\mu = 1$ is equivalent to

(i) $P(\mathbf{s}(\mathcal{U}_n) \in \bar{\mathbf{T}}_{[I]} \setminus \mathbf{T}_{[I]}) \to 0$ and

(ii) $(n^{-1/2}l^*(\mathbf{s}(\mathcal{U}_n)), n \to \infty)$ is stochastically bounded above, and $(n^{-1/2}l_*(\mathbf{s}(\mathcal{U}_n)), n \to \infty)$ is stochastically bounded below, where $l^*(\mathbf{s})$ and $l_*(\mathbf{s})$ denote the longest and shortest edge-lengths of $\mathbf{s}$.

It would be possible to verify (i) by modifying a similar argument in [6], and to verify (ii) by estimating tails in (10) – but the details are messy. Instead, we give an analytic verification that $\mu = 1$ in section 4.4. This establishes Theorem 1 along with properties (i) and (ii). To prove Theorem 2, for $V \subset [I]$ let $\mathbf{s}_V(\mathcal{U}_n)$ be the subtree of $\mathcal{U}_n$ spanned by V. The argument above represents $\mathcal{S}_\mathbf{c}$ as a weak limit of $\mathbf{s}(\mathcal{U}_n)$ with rescaled edge-lengths, which implies that the spanning subtree $\mathcal{S}_\mathbf{c}^V$ appears as the weak limit of $\mathbf{s}_V(\mathcal{U}_n)$ with rescaled edge-lengths. Repeating the proof of Proposition 6 for $\mathbf{s}_V(\mathcal{U}_n)$ in place of $\mathbf{s}(\mathcal{U}_n)$ yields the following.

For each H with $V \subseteq H \subseteq [I]$ and each $\mathbf{t} \in T_H$ such that $\mathbf{t}$ is spanned by V; for each $\boldsymbol{l}^* = (l_e^*, e \in \mathcal{E}(\mathbf{t}))$ with $s = \sum_e l_e^*$, and each $n^{-1/2}\boldsymbol{l} \to \boldsymbol{l}^*$;

$$\begin{aligned} & P\left(\text{hubs}(\mathbf{s}_V(\mathcal{U}_n)) = H, \text{shape}(\mathbf{s}_V(\mathcal{U}_n)) = \mathbf{t}, \text{lengths}(\mathbf{s}_V(\mathcal{U}_n)) = \boldsymbol{l}\right) \\ & = \left(\prod_{h \in H} c_h^{D_h \mathbf{t} - 1} \quad (s + c_H) \exp(-\tfrac{1}{2}s^2 - sc) \; + o(1) \right) \; n^{-\#\mathcal{E}(\mathbf{t})/2}. \end{aligned}$$

So to deduce the equality in (3) it suffices to show that no mass is lost in the limit, i.e. to establish the analogs of (i) and (ii) for $(\mathbf{s}_V(\mathcal{U}_n))$. But these are immediate consequences of (i) and (ii) for $(\mathbf{s}(\mathcal{U}_n))$. □

4 Distributions associated with $\mathcal{S}_\mathbf{c}$

4.1 Distribution of the shape

For $m = 1, 2, \ldots$ and $x > 0$ define

$$\Psi_m(x) := \int_0^\infty s^{m-1} e^{-\frac{1}{2}s^2 - sx} ds \tag{11}$$

and note the recursion

$$\Psi_{m+1}(x) + x\Psi_m(x) = (m-1)\Psi_{m-1}(x) \qquad (m > 1) \tag{12}$$

obtained via integration by parts. The function $\Psi_m(x)$ is a variation of the repeated integral of the error function, with well known expressions in terms of parabolic cylinder functions or the confluent hypergeometric function [1, 7.2].

Proposition 7 *If $I \geq 3$ then for each $\mathbf{t} \in T_{[I]}$ with m edges,*

$$P(\text{shape}(\mathcal{S}_\mathbf{c}) = \mathbf{t}) = \left(\prod_{i=1}^I c_i^{D_i \mathbf{t} - 1} \right) \frac{\Psi_{m-1}(c)}{(m-2)!}.$$

Proof. Consider $\mathbf{t} \in T_{[I]}$ with $\#\mathcal{E}(\mathbf{t}) = m$. By integration of the basic formula (1) over all length vectors with total length s,

$$P(\text{shape}(\mathcal{S}_\mathbf{c}) = \mathbf{t}, L(\mathcal{S}_\mathbf{c}) \in ds) = \left(\prod_{i=1}^I c_i^{D_i \mathbf{t} - 1} \right) \frac{s^{m-1}}{(m-1)!} (s + c) e^{-\frac{1}{2}s^2 - sc} ds. \tag{13}$$

Integrating out s and applying the recursion (12) gives the stated formula. □

4.2 Enumeration of spanning trees by degree sequence

We start with an enumeration which is the basis of all subsequent calculations.

Proposition 8 *For each $0 \le d \le I-2$ and each vector of non-negative integers $(d_i, 1 \le i \le I)$ with $\sum_i d_i = d$, let $T(d_1, \ldots, d_I)$ be the set of trees $\mathbf{t} \in T_{[I]}$ such that $D_i\mathbf{t} - 1 = d_i$ for all $1 \le i \le I$. Then*

$$\#T(d_1, \ldots, d_I) = \binom{d}{d_1, \ldots, d_I} \frac{(2I-d-4)!}{d!(I-d-2)!} 2^{d-I+2}. \tag{14}$$

Proof. According to *Cayley's multinomial theorem* [7, 21, 16], for $d = I-2$ the multinomial coefficient in (14) is the number of trees labeled by $[I]$ in which the excess degree of vertex i is d_i for each $i \in [I]$. For a tree $\mathbf{t} \in T(d_1, \ldots, d_I)$ let u be the number of unlabeled vertices of $\mathbf{t}$. By (17), $u = I - d - 2$. Let $\hat{T}(d_1, \ldots, d_I)$ be the set of all trees labeled by $[I+u]$ in which vertex i has excess degree d_i for $1 \le i \le I$ and vertex $I+j$ has excess degree 2 for $1 \le j \le u$. By Cayley's multinomial theorem,

$$\#\hat{T}(d_1, \ldots, d_I) = \binom{d+2u}{d_1, \ldots, d_I, 2, \ldots 2} = \frac{(2I-d-4)!}{d_1! \ldots d_I! 2^{I-d-2}}. \tag{15}$$

Because the unlabeled vertices of a tree $\mathbf{t} \in T_{[I]}$ are implicitly labeled by their locations in $\mathbf{t}$ relative to the vertices in I, the delabeling map from $\hat{T}(d_1, \ldots, d_I)$ to $T(d_1, \ldots, d_I)$ is $u!$ to 1, and the equality (14) follows by dividing both sides of (15) by $u!$. □

For a tree $\mathbf{t} \in T_{[I]}$, let

$$D(\mathbf{t}) := \sum_{i=1}^{I} (D_i\mathbf{t} - 1)$$

and call $D(\mathbf{t})$ the *excess degree* of $\mathbf{t}$. The set of possible values of the excess degree is $\{0, 1, \ldots, I-2\}$. It is easy to check the counting formulas

$$\#\mathcal{E}(\mathbf{t}) = 2I - 3 - D(\mathbf{t}) \tag{16}$$

$$\#(\text{unlabeled vertices of } \mathbf{t}) = I - D(\mathbf{t}) - 2. \tag{17}$$

Sum formula (14) over all $(d_i, 1 \le i \le I)$ with $\sum_i d_i = d$ and use the multinomial theorem to see that

$$\#T_{[I]} = \sum_{d=0}^{I-2} I^d \, \frac{(2I-d-4)! 2^{d-I+2}}{d!(I-d-2)!} \tag{18}$$

where the dth term is the number of $\mathbf{t} \in T_{[I]}$ with excess degree d.

4.3 Distribution of the total length

For a tree with edge lengths $\mathbf{t}$, write

$$L(\mathbf{t}) := \sum_{e \in \mathcal{E}(\mathbf{t})} l_e$$

for the total edge-length of $\mathbf{t}$. The density of $L(\mathcal{S}_\mathbf{c})$ induced by formula (1) will now be derived. It will be checked in the next section that this density integrates to 1 for all choices of I and $\mathbf{c}$. This constitutes a proof of Theorem 1, as the only point in doubt is the value of a normalization constant. Let $(d_i, 1 \le i \le I)$ be non-negative integers with $\sum_i d_i = d \le I-2$. Formula (13) and Proposition 8 imply that for $s > 0$

$$P(D_i(\mathcal{S}_\mathbf{c}) - 1 = d_i, 1 \le i \le I; L(\mathcal{S}_\mathbf{c}) \in ds)/ds =$$

$$\binom{d}{d_1, \ldots, d_I} \left(\prod_{i=1}^{I} c_i^{D_i \mathbf{t} - 1} \right) \frac{1}{(I-2)!} \binom{I-2}{d} 2^{d-I+2}\, s^{2I-d-4} (s+c) e^{-\frac{1}{2}s^2 - sc}. \qquad (19)$$

So the multinomial theorem gives

$$P(D(\mathcal{S}_\mathbf{c}) = d, L(\mathcal{S}_\mathbf{c}) \in ds)/ds = \frac{c^d 2^{d-I+2}}{(I-2)!} \binom{I-2}{d} s^{2I-4-d} (s+c) e^{-\frac{1}{2}s^2 - sc}. \qquad (20)$$

Note that this joint distribution depends only on I and $c := \sum_{i=1}^{I} c_i$. Now sum over d to deduce the formula for the density of $L(\mathcal{S}_\mathbf{c})$ stated in part (i) of the following corollary of Theorem 1. The remaining parts of the corollary then follow easily. The corollary shows how to construct a random tree $\mathcal{S}_\mathbf{c}$ with the distribution defined by the basic formula by a five step process from more elementary ingredients. For instance, for modest values of I it would be quite feasible to simulate $\mathcal{S}_\mathbf{c}$ by computer using this construction.

Corollary 9 *For $I \ge 2$ and $\mathbf{c} := (c_1, \ldots, c_I)$ with $c_i \ge 0$ and $\sum_i c_i = c$*
(i) *the density of $L(\mathcal{S}_\mathbf{c})$ at $s > 0$ is*

$$P(L(\mathcal{S}_\mathbf{c}) \in ds)/ds = \frac{1}{(I-2)!} \left(\frac{s}{2}\right)^{I-2} (s+2c)^{I-2} (s+c) e^{-\frac{1}{2}s^2 - sc}; \qquad (21)$$

(ii) *the law of $D(\mathcal{S}_\mathbf{c})$ given $L(\mathcal{S}_\mathbf{c}) = s$ is binomial $(I-2, 2c/(s+2c))$:*

$$P(D(\mathcal{S}_\mathbf{c}) = d \,|\, L(\mathcal{S}_\mathbf{c}) = s) = \binom{I-2}{d} \left(\frac{2c}{s+2c}\right)^d \left(\frac{s}{s+2c}\right)^{I-2-d}; \qquad (22)$$

(iii) *for $c > 0$, given $L(\mathcal{S}_{\mathbf{c}}) = s$ and $D(\mathcal{S}_{\mathbf{c}}) = d$, the joint law of the $(D_i(\mathcal{S}_{\mathbf{c}})-1, 1 \le i \le I)$ is multinomial with parameters d and $(c_i/c, 1 \le i \le I)$;*
(iv) *given $L(\mathcal{S}_{\mathbf{c}}) = s$ and $D_i - 1 = d_i$ for $1 \le i \le I$, the shape of $\mathcal{S}_{\mathbf{c}}$ is picked uniformly at random from the set $T(d_1, \ldots, d_I)$ of all trees in T_I with the given excess degree sequence, as enumerated in (14);*
(v) *given $L(\mathcal{S}_{\mathbf{c}}) = s$ and that the shape of $\mathcal{S}_{\mathbf{c}}$ equals $\mathbf{t}$ with $m := 2I - 3 - d$ edges, the m segment lengths of $\mathcal{S}_{\mathbf{c}}$ are distributed as the spacings between $m-1$ independent uniform $(0, s)$ variables.*

From (16) and (ii) above, for the number $\#\mathcal{E}(\mathcal{S}_{\mathbf{c}})$ of edges of $\mathcal{S}_{\mathbf{c}}$, we find that

the distribution of $\#\mathcal{E}(\mathcal{S}_{\mathbf{c}}) - I + 1$ given $L(\mathcal{S}_{\mathbf{c}}) = s$ is binomial $(I-2, s/(s+2c))$.

Also, (ii) and (iii) combine to show that

> *the conditional distribution of $(I - 2 - D; D_i, i \in [I])$ given $L(\mathcal{S}_{\mathbf{c}}) = s$ is multinomial with parameters $I-2$ and $(s/(s+2c); 2c_i/(s+2c), i \in [I])$.*

Several further implications of the corollary are spelled out in following sections.

4.4 Checking the constant of integration

Writing $f_{I,c}(s)$ for the right side of (21), it suffices to verify that for fixed $c \ge 0$

$$\int_0^\infty f_{I,c}(s)ds = 1, \ 2 \le I < \infty. \tag{23}$$

But for $0 \le z < 1$ we can compute

$$\sum_{I=0}^\infty z^I \int_0^\infty f_{I+2,c}(s)ds \qquad \int_0^\infty \sum_{I=0}^\infty \frac{z^I}{I!}\left(\frac{s}{2}\right)^I (s+2c)^I (s+c) e^{-\frac{1}{2}s^2 - sc}\, ds$$
$$= \int_0^\infty (s+c) \exp\left(-(1-z)(\tfrac{1}{2}s^2 + sc)\right)\, ds = \frac{1}{1-z}$$

and (23) follows.

4.5 Distribution of the excess degree

Corollary 9 specified the distribution of the length of $\mathcal{S}_{\mathbf{c}}$ and the conditional law of the excess degrees $D_i(\mathcal{S}_{\mathbf{c}})$ given the length. Integrating out the length yields the two formulae stated in the following proposition, which can also be deduced from Propositions 8 and 7. Recall that $D(\mathcal{S}_{\mathbf{c}}) := \sum_{i=1}^I D_i(\mathcal{S}_{\mathbf{c}})$ and that this number determines both the number of edges of $\mathcal{S}_{\mathbf{c}}$ and the number of unlabeled vertices of $\mathcal{S}_{\mathbf{c}}$ via (16) and (17). So the distribution of either of these numbers can be read from that of $D(\mathcal{S}_{\mathbf{c}})$.

Proposition 10 *Let $I \geq 3$.*
(i) *For each $0 \leq d \leq I-2$, and each vector of non-negative integers $(d_i, 1 \leq i \leq I)$ with $\sum_i d_i = d$,*

$$P(D_i(\mathcal{S}_{\mathbf{c}}) - 1 = d_i,\ 1 \leq i \leq I) = \binom{d}{d_1, \ldots, d_I} \frac{(2I-d-4)2^{d-I+2}}{d!(I-d-2)!} \Psi_{2I-d-4}(c) \prod_{i=i}^{I} c_i^{d_i}.$$

(ii)

$$P(D(\mathcal{S}_{\mathbf{c}}) = d) = \frac{(2I-d-4)2^{d-I+2}}{d!(I-d-2)!}\, c^d\, \Psi_{2I-d-4}(c), \quad 0 \leq d \leq I-2. \tag{24}$$

Note the implication of (ii) that the distribution of $D(\mathcal{S}_{\mathbf{c}})$ depends only on the sum c of $\mathbf{c}$. The consequence of (ii), that the right side of (24) sums to 1 as d ranges from 0 to $I-2$, can also be checked using the recursion (12).

4.6 A coincidence in distribution

There is a remarkable coincidence between the distribution of $D(\mathcal{S}_{\mathbf{c}})$ displayed in Proposition 10, and a distribution derived from sampling the excursion intervals of a Brownian motion $B := (B_t, 0 \leq t \leq 1)$. Let $(L_t, t \geq 0)$ be the usual local time process of B at 0. Let K_n be the number of equivalence classes of the random partition of $[n]$ defined by the random equivalence relation $i \sim j$ iff there is no zero of B between times U_i and U_j, where the U_i are i.i.d. uniform $[0,1]$ random variables independent of B. As observed in [18], $L_1 = \lim_{n\to\infty} K_n/\sqrt{2n}$ almost surely. It can be deduced from results of [15, 18] that for each $n = 2, 3, \ldots$ the conditional distribution of K_n given $L_1 = c$ is identical to the distribution of $D(\mathcal{S}_{\mathbf{c}}) + 1$ as determined by (24) for any $\mathbf{c} = (c_1, \ldots, c_I)$ with $I = n+1$ and $\sum_{i=1}^{I} c_i = c$. Due to results of [4], this distribution of K_n given $L_1 = c$ can also be interpreted as the distribution of number of components of the partition of $[n]$ generated as follows: first construct a Brownian CRT, then pick n points $X_1, \ldots, X_n$ uniformly at random from the mass measure of the CRT, and partition $[n]$ by the random equivalence relation $i \sim j$ iff the path from X_i to X_j in the CRT contains no point of a Poisson process of rate c per unit length on the skeleton of the tree. As explained in section 7, the subtree of the Brownian CRT spanned by $X_i, 1 \leq i \leq n$ is a copy of $\mathcal{S}_{\mathbf{c}}$ for $\mathbf{c}$ a vector of zeros of length n. These observations can be developed to give an essentially combinatorial proof of the coincidence in distribution between K_n given $L_1 = c$ and $D(\mathcal{S}_{\mathbf{c}}) + 1$ for $\mathbf{c} = (c_1, \ldots, c_I)$ with $I = n+1$, but the argument is tricky and will not be attempted here.

5 Scaling and limiting cases of c

5.1 Some specific limits

For a tree $\mathbf{s}$ with edge-lengths (l_e) and for $0 < a < \infty$ write $a \otimes \mathbf{s}$ for the tree whose edge-lengths are (al_e). In this section we shall see that in several limit cases rescaled edge-lengths become i.i.d. with exponential distribution.

Case 1. Consider $\mathbf{c} = (c_1, \ldots, c_I) = (\alpha, 0, 0, \ldots, 0)$. Write $\mathbf{star}_I$ for the discrete tree in which each vertex $2 \le i \le I$ is connected to vertex 1 by an edge. Then as $\alpha \to \infty$ there is the convergence in distribution

$$\alpha \otimes \mathcal{S}_{\mathbf{c}} \stackrel{d}{\to} \mathbf{star}_I(\text{ with independent exponential(1) edge-lengths}). \tag{25}$$

To see why, the basic formula implies

$$P(\text{shape}(\mathcal{S}_{\mathbf{c}}) = \mathbf{star}_I, \text{lengths}(\mathcal{S}_{\mathbf{c}}) \in [\boldsymbol{l}, \boldsymbol{l} + d\boldsymbol{l}])/d\boldsymbol{l} = \alpha^{I-2}(\alpha + s)\exp(-\tfrac{1}{2}s^2 - s\alpha)$$

where $s = \sum_e l_e$. Multiplying edge-lengths by α gives

$$P(\text{shape}(\mathcal{S}_{\mathbf{c}}) = \mathbf{star}_I, \alpha \times \text{lengths}(\mathcal{S}_{\mathbf{c}}) \in [\boldsymbol{l}, \boldsymbol{l} + d\boldsymbol{l}])/d\boldsymbol{l} = (1 + \tfrac{s}{\alpha^2})\exp(-\tfrac{s^2}{2\alpha^2} - s)$$

As $\alpha \to \infty$ this density tends to e^{-s}, which is the joint density of $I-1$ i.i.d. exponential(1) variables $(\eta_e, e \in \mathcal{E}(\mathbf{star}_I))$.

Case 2. $\mathbf{c} = (c_1, \ldots, c_I) = (\alpha, \alpha, \alpha, \ldots, \alpha)$. Write $\mathbf{u}_I$ for the random tree with edge lengths obtained by first picking a discrete tree $\mathbf{t} \in U_{[I]}$ uniformly from all I^{I-2} trees in $U_{[I]}$ and then making the edge-lengths be independent exponential(1). Then

$$I\alpha \otimes \mathcal{S}_{\mathbf{c}} \stackrel{d}{\to} \mathbf{u}_I \text{ as } \alpha \to \infty. \tag{26}$$

To see why, for $\mathbf{t} \in U_{[I]}$ the basic formula implies

$$P(\text{shape}(\mathcal{S}_{\mathbf{c}}) = \mathbf{t}, \text{lengths}(\mathcal{S}_{\mathbf{c}}) \in [\boldsymbol{l}, \boldsymbol{l} + d\boldsymbol{l}])/d\boldsymbol{l} = \alpha^{I-2}(I\alpha + s)\exp(-\tfrac{1}{2}s^2 - sI\alpha)$$

where $s = \sum_e l_e$. Multiplying edge-lengths by $I\alpha$ gives

$$\begin{aligned} & P(\text{shape}(\mathcal{S}_{\mathbf{c}}) = \mathbf{t}, I\alpha \times \text{lengths}(\mathcal{S}_{\mathbf{c}}) \in [\boldsymbol{l}, \boldsymbol{l} + d\boldsymbol{l}])/d\boldsymbol{l} \\ = \;& \tfrac{1}{I^{I-2}}(1 + \tfrac{s}{I^2\alpha^2})\exp(-\tfrac{s^2}{2I^2\alpha^2} - s) \\ \to \;& \tfrac{1}{I^{I-2}}e^{-s} \\ = \;& P(\text{shape}(\mathbf{u}_I) = \mathbf{t}, \text{lengths}(\mathbf{u}_I) \in [\boldsymbol{l}, \boldsymbol{l} + d\boldsymbol{l}])/d\boldsymbol{l}. \end{aligned}$$

Case 3. $\mathbf{c} = (c_1, \ldots, c_I) = (0, 0, 0, \ldots, 0)$. By (2) and symmetry in the basic formula, shape$(\mathcal{S}_\mathbf{c})$ is uniform on the subset $T^0_{[I]} \subset T_{[I]}$ of discrete trees in which each labeled vertex has degree 1. It is well known (and a special case of Proposition 8) that

$$\#T^0_{[I]} = \frac{(2I-4)!}{(I-2)!2^{I-2}} = (2I-5) \times (2I-7) \times \ldots \times 3 \times 1.$$

Using Corollary 9(i),

$$P(L(\mathcal{S}_\mathbf{c}) \in ds)/ds = \frac{1}{(I-2)!} \left(\frac{s^2}{2}\right)^{I-2} s e^{-s^2/2}.$$

It follows that $L(\mathcal{S}_\mathbf{c}) \sim \sqrt{2I}$ as $I \to \infty$. From Corollary 9(v) and routine properties of spacings, for fixed k the joint distribution $(l_{e_1}, \ldots, l_{e_k})$ of any k of the $2I$ edge-lengths satisfies: as $I \to \infty$

$$\sqrt{2I}\ (l_{e_1}, \ldots, l_{e_k}) \xrightarrow{d} \text{ independent exponential(1).}$$

5.2 Limits of degree distributions

Recall that the excess degree $D(\mathcal{S}_\mathbf{c})$ is between 0 and $I-2$. The next result, which follows by routine arguments from the exact formulas in Corollary 9, indicates the asymptotic regime in which the excess degree is between these extremes.

Corollary 11 *Consider a sequence of vectors* $\mathbf{c} = (c_1, \ldots, c_I)$, $c = \sum_i c_i$, *such that*

$$I \to \infty; \quad c/\sqrt{2I} \to \alpha \in [0, \infty].$$

Then
(i) $L(\mathcal{S}_\mathbf{c})/\sqrt{2I} \xrightarrow{p} \sqrt{\alpha^2+1} - \alpha$.
(ii) $\frac{D(\mathcal{S}_\mathbf{c})}{I} \xrightarrow{p} \frac{2\alpha}{\sqrt{\alpha^2+1}+\alpha}$.
(iii) If also $c_i \sim \lambda/\sqrt{2I}$ *for some* $0 \le \lambda \le \infty$ *then*

$$D_i(\mathcal{S}_\mathbf{c}) - 1 \xrightarrow{d} \text{Poisson}\left(\frac{\lambda}{\sqrt{\alpha^2+1}+\alpha}\right)$$

interpreting the limit as ∞ *when* $\lambda = \infty$.

6 Applications of the master formula

Throughout section 6, $\mathbf{c} := (c_1, \ldots, c_I) \in \mathcal{C}$ is fixed and $\sum_i c_i = c$. By application of Proposition 8 and appropriate summations and integrations, using the master formula in place of the basic formula, there is the following analog of (19):

Proposition 12 *Let $V \subseteq [I]$ have $\#V \geq 2$. Let $\mathcal{S}_{\mathbf{c}}^V$ be the subtree of $\mathcal{S}_{\mathbf{c}}$ spanned by V. Then for each H with $V \subseteq H \subseteq I$, each possible excess degree sequence $(d_h, h \in H)$ of non-negative integers with $\sum_h d_h = d$, where $0 \leq d \leq \#H - 2$, and each $s > 0$*

$$P(\mathrm{hubs}(\mathcal{S}_{\mathbf{c}}^V) = H, D_h(\mathcal{S}_{\mathbf{c}}^V) - 1 = d_h \text{ for } h \in H, L(\mathcal{S}_{\mathbf{c}}^V) \in ds)/ds$$

$$= \frac{s^{2\#H-4-d} \left(\prod_{h\in H} c_h^{d_h}\right) (s + c_H) e^{-\frac{1}{2}s^2 - sc}}{\left(\prod_{h\in H} d_h!\right) (\#H - d - 2)! 2^{\#H-d-2}} \tag{27}$$

Proof of Corollary 3. It is enough to consider the case $i = 1$ and $j = 2$. When $V = [2]$, the spanning subtree $\mathcal{S}_{\mathbf{c}}^{[2]}$ consists of a path from vertex 1 to vertex 2 passing through some set $A \subseteq [I] \setminus [2]$ of other vertices. So (27) gives

$$P(\mathrm{hubs}(\mathcal{S}_{\mathbf{c}}^{[2]}) = [2] \cup A, L(\mathcal{S}_{\mathbf{c}}^{[2]}) \in ds)/ds = s^{\#A} \Pi_A (s + c_1 + c_2 + c_A) e^{-\frac{1}{2}s^2 - sc}$$

where $\Pi_A := \prod_{i\in A} c_i$ and $c_A := \sum_{i\in A} c_i$. Summing over all $A \subseteq [I] \setminus [2]$ gives a formula for the density of $L(\mathcal{S}_{\mathbf{c}}^{[2]})$ at s which can be simplified by application of the following elementary identities of polynomials in variables $x_b, b \in B$ applied to $B = [I] \setminus [2]$:

$$\sum_{A\subseteq B} \prod_{a\in A} x_a = \prod_{b\in B} (1 + x_b);$$

$$\sum_{A\subseteq B} \left(\prod_{a\in A} x_a\right) \left(\sum_{b\in A} x_b\right) = \left(\sum_{a\in B} \frac{x_a^2}{1 + x_a}\right) \prod_{b\in B} (1 + x_b).$$

The result of this simpification is

$$P((L(\mathcal{S}_{\mathbf{c}}^{[2]}) \in ds)/ds = \left(s + c_1 + c_2 + \sum_{i=3}^{I} \frac{c_i^2 t}{1 + c_i t}\right) e^{-\frac{1}{2}s^2 - sc} \prod_{j=3}^{I} (1 + c_j s) \tag{28}$$

$$= -\frac{d}{ds} \left(e^{-\frac{1}{2}s^2 - sc} \prod_{j=3}^{I} (1 + c_j s)\right)$$

which yields Corollary 3. □

As will be described in section 7.4, there is some motivation for studying the length of the subtree $\mathcal{S}_{\mathbf{c}}^{[k]}$ when $c_1 = c_2 = \ldots = c_k = 0$ but $c_i > 0$ for $k < i \leq I$. For $k = 2$ this is the special case $c_1 = c_2 = 0$ of Corollary 3. In principle we can derive the length distribution from the master formula for general k. But the result is complicated, so we record only the case $k = 3$. The subtree spanned by $\{1, 2, 3\}$ must have three edges meeting at a vertex of degree 3, which might be hub i for some $i > 3$, or an unlabeled junction point. Applying the master formula to each possibility yields the following conclusion:

Corollary 13 *If* $c_1 = c_2 = c_3 = 0$ *then*

$$P(L(\mathcal{S}_{\mathbf{c}}^{[3]}) \in ds)/ds = \sum_{A \subseteq [I] \setminus [3]} \tfrac{1}{2} s^{\#A+1} \Pi_A (s + c_A)^2 e^{-\frac{1}{2}s^2 - sc}$$

where the Ath term equals $P(\text{hubs}(\mathcal{S}_{\mathbf{c}}^{[3]}) = [3] \cup A)$.

7 Interpretation as spanning subtrees in the ICRT

7.1 Some abstract theory

We first outline very briefly some abstract theory. Let $v_1, \ldots, v_n$ be a uniform random ordering of the vertices of some n-vertex random tree with edge lengths. For $2 \leq k \leq n$ let $\mathcal{R}_k$ be the subtree spanned by $\{v_1, \ldots, v_k\}$, with these vertices relabeled as $\{1+, 2+, \ldots, k+\}$ and other vertices unlabeled. Then the family $(\mathcal{R}_k; 2 \leq k \leq n)$ automatically has the properties
(i) The distribution of $\mathcal{R}_k$ is invariant under permutations of the labels $\{1+, 2+, \ldots, k+\}$.
(ii) $\mathcal{R}_k$ is distributed as the subtree of $\mathcal{R}_{k+1}$ spanned by $\{1+, 2+, \ldots, k+\}$.
Now suppose we are given an infinite family $(\mathcal{R}_k; 2 \leq k < \infty)$ satisfying (i) and (ii), and such that each vertex $j+$ is a leaf of $\mathcal{R}_k$ for $k \geq j$. Under extra technical conditions, [3] Theorem 3 asserts there exists a representing *continuum random tree* $\mathcal{T}$. Roughly, a realization of $\mathcal{T}$ is a tree with edge lengths with an uncountable set of leaves, and with a non-atomic probability measure μ on the leaves. One can therefore pick (conditionally on $\mathcal{T}$ and μ) independent leaves $v_1, v_2, \ldots$ with distribution μ, and the "representation" is that

$$\mathcal{R}_k \text{ is distributed as the subtree of } \mathcal{T} \text{ spanned by } \{v_1, \ldots, v_k\}. \tag{29}$$

To see the relevance of this abstract theory to our model, consider $\mathbf{c} = (c_1, \ldots, c_I)$ where $0 \leq I < \infty$ and $c_i > 0$ for $1 \leq i \leq I$. For $k \geq 0$ write $\mathbf{c}[+k]$ for the vector obtained

by appending k zero terms to $\mathbf{c}$:

$$\begin{aligned}(\mathbf{c}[+k])_i &= c_i, \quad i \le I \\ &= 0, \quad I+1 \le i \le I+k.\end{aligned} \tag{30}$$

To avoid trivialities, suppose $I + k \ge 2$. In the associated tree $\mathcal{S}_{\mathbf{c}[+k]}$ relabel vertices $I+1, \ldots, I+k$ as $1+, \ldots, k+$. Write $\mathcal{R}^{\mathbf{c}}_k$ for the subtree of $\mathcal{S}_{\mathbf{c}[+k]}$ spanned by $\{1+, \ldots, k+\}$. The vertices $j+$ are leaves by (2). The family $(\mathcal{R}^{\mathbf{c}}_k; 2 \le k < \infty)$ satisfies (i) because, by symmetry in the basic formula, the distribution of $\mathcal{S}_{\mathbf{c}[+k]}$ is invariant under permutations of the labels $\{1+, 2+, \ldots, k+\}$. Similarly, to check (ii) it is enough to check $\mathcal{S}_{\mathbf{c}[+k]}$ is distributed as the subtree of $\mathcal{S}_{\mathbf{c}[+k+1]}$ spanned by $\{1+, 2+, \ldots, k+\}$. But this follows from the master formula, since the hubs of this subtree are evidently $[I] \cup \{1+, \ldots, k+\}$.

Thus by checking the technical conditions in [3] one could establish the existence of a representing continuum random tree, say $\widehat{\mathcal{T}}^{\mathbf{c}}$. However, Aldous and Pitman [2] give a more algorithmic construction (reviewed in section 7.2) of an *inhomogeneous continuum random tree* (ICRT), which we shall see (Proposition 14) is the same object up to parametrization. As described in section 7.4, the problem studied in [2] motivates some difficult distributional questions concerning $\mathcal{S}_{\mathbf{c}}$.

7.2 The line-breaking construction

This construction is from [2] section 2. Fix $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_I)$ with $0 \le I < \infty$, with each $\theta_i > 0$ and such that $\sum_i \theta_i^2 \le 1$. Define $a = 1 - \sum_i \theta_i^2$. If $a > 0$ let $((U_j, V_j), 1 \le j < \infty)$ with $0 < U_1 < U_2 < \ldots$ be the points of a Poisson point process of rate a per unit area on the octant $\{(u, v) : 0 < v < u < \infty\}$. For each $i \ge 1$ such that $\theta_i > 0$, let $0 < \xi_{i,1} < \xi_{i,2} < \ldots$ be the points of a Poisson point process on $(0, \infty)$ of rate θ_i per unit length. These are the "random" ingredients of our construction. The construction is illustrated in figures 2,3 and 4. In outline, we cut the line $[0, \infty)$ into finite-length segments and reassemble the segments as "branches" of a tree, where each point of the tree is labeled by some $0 \le x < \infty$, and then pass to a completion. Here are the details.

Call each point U_j a 0-*cutpoint*, and say that V_j is the corresponding *joinpoint*. Call each point $\xi_{i,j}$ with $\theta_i > 0$ and $j \ge 2$ (note the 2) an i-*cutpoint*, and say that $\xi_{i,1}$ is the corresponding *joinpoint*. Note that there are (with probability 1, a qualification in effect throughout the construction) only finitely many cutpoints in any finite interval $[0, x]$, because for $i \ge 1$ the mean number of i-cutpoints in that interval equals $\theta_i x - (1 - \exp(-\theta_i x)) \le \theta_i^2 x^2$. We may therefore order the cutpoints as $0 < \eta_1 < \eta_2 < \ldots$, where $\eta_k \to \infty$ as $k \to \infty$. Figure 2 illustrates the cutpoints, with each η_k identified as some U_j or $\xi_{i,j}$.

We build the tree by starting with the branch $[0, \eta_1]$ and then, inductively on $k \geq 1$, attaching the branch $(\eta_k, \eta_{k+1}]$ to the joinpoint η_k^* corresponding to the cutpoint η_k. Figure 3 illustrates the attachment of the first 8 branches, using the realization in figure 2. The reader will find it helpful to work through the construction in figure 3: the sequence of attachments of branches is

$$[0, U_1], \quad (V_1, U_2], \quad (V_2, \xi_{1,2}], \quad (\xi_{1,1}, \xi_{4,2}], \quad (\xi_{4,1}, U_3], \quad (V_3, \xi_{2,2}], \quad (\xi_{2,1}, \xi_{1,3}], \quad (\xi_{1,1}, U_4].$$

In [2] the emphasis was on continuing this construction over the infinite line $[0, \infty)$ to yield a realization of an ICRT $\mathcal{T}_{\boldsymbol{\theta}}$, but for the purposes of this paper we need only consider finite numbers of branches. Given $\boldsymbol{\theta}$ and $k \geq 0$, stop the construction at the first cutpoint η_J such that $J \geq \max(1, k-1)$ and such that the interval $[0, \eta_J]$ contains each $\xi_{i,1}, 1 \leq i \leq I$. This gives a tree with edge lengths, as in figure 3. For each $1 \leq i \leq I$, relabel the point $\xi_{i,1}$ as *hub* i. And for each $1 \leq j \leq J$ relabel point η_{j-1} as leaf $j+$ (take $\eta_0 = 0$). This yields a tree with edge lengths (see figure 4) with I hubs and with some number $J \geq k$ of leaves $j+$ which span the tree. Finally, define $\tilde{\mathcal{S}}_{\boldsymbol{\theta}[+k]}$ to be the subtree spanned by the hubs $[I]$ and the subset of leaves $\{1+, \ldots, k+\}$, and define $\tilde{\mathcal{S}}_{\boldsymbol{\theta}}$ to be the subtree spanned by the hubs $[I]$ only.

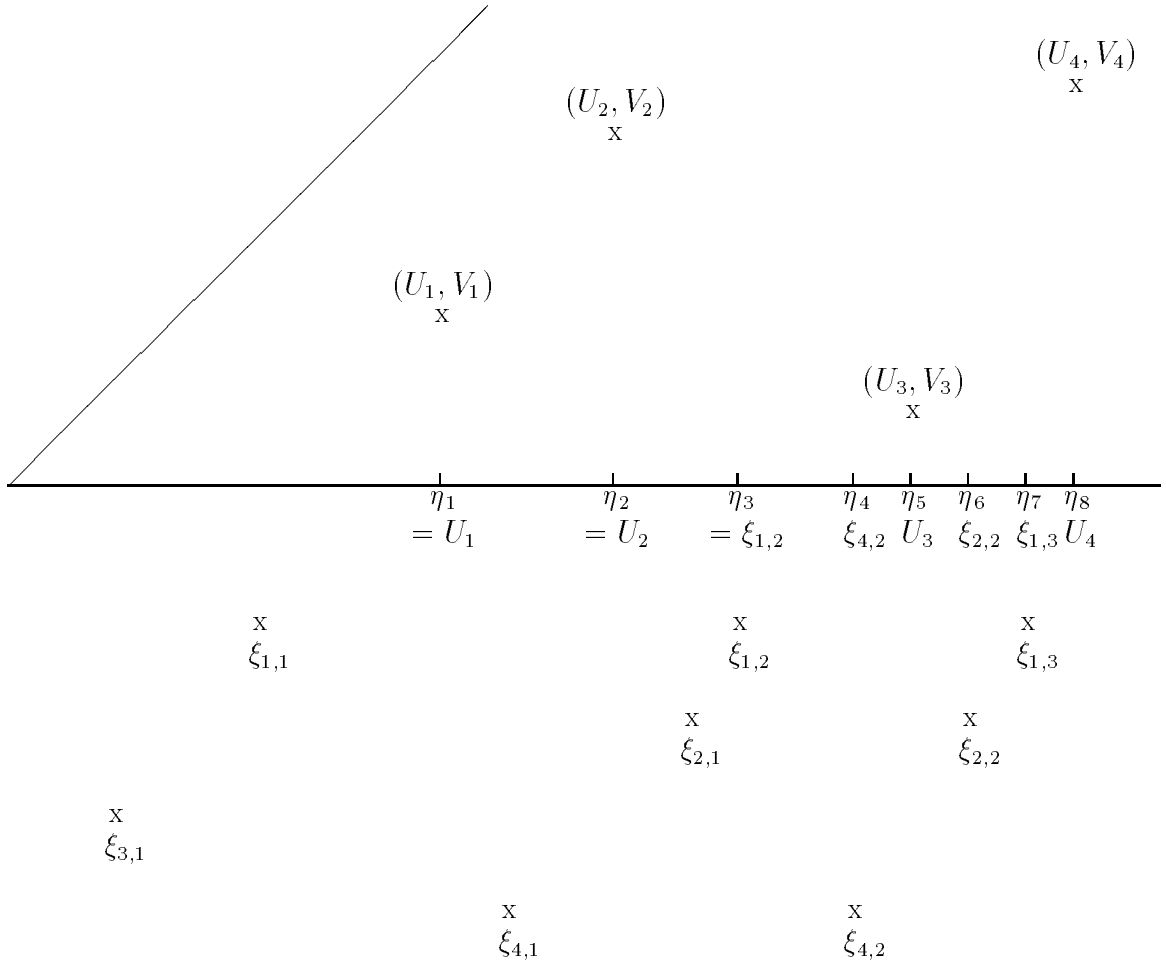
(U_4, V_4)
(U_2, V_2)
(U_1, V_1)
(U_3, V_3)
$\eta_1 = U_1$
$\eta_2 = U_2$
$\eta_3 = \xi_{1,2}$
$\eta_4 \; \xi_{4,2}$
$\eta_5 \; U_3$
$\eta_6 \; \xi_{2,2}$
$\eta_7 \; \xi_{1,3}$
$\eta_8 \; U_4$
$\xi_{1,1}$
$\xi_{1,2}$
$\xi_{1,3}$
$\xi_{2,1}$
$\xi_{2,2}$
$\xi_{3,1}$
$\xi_{4,1}$
$\xi_{4,2}$

Figure 2

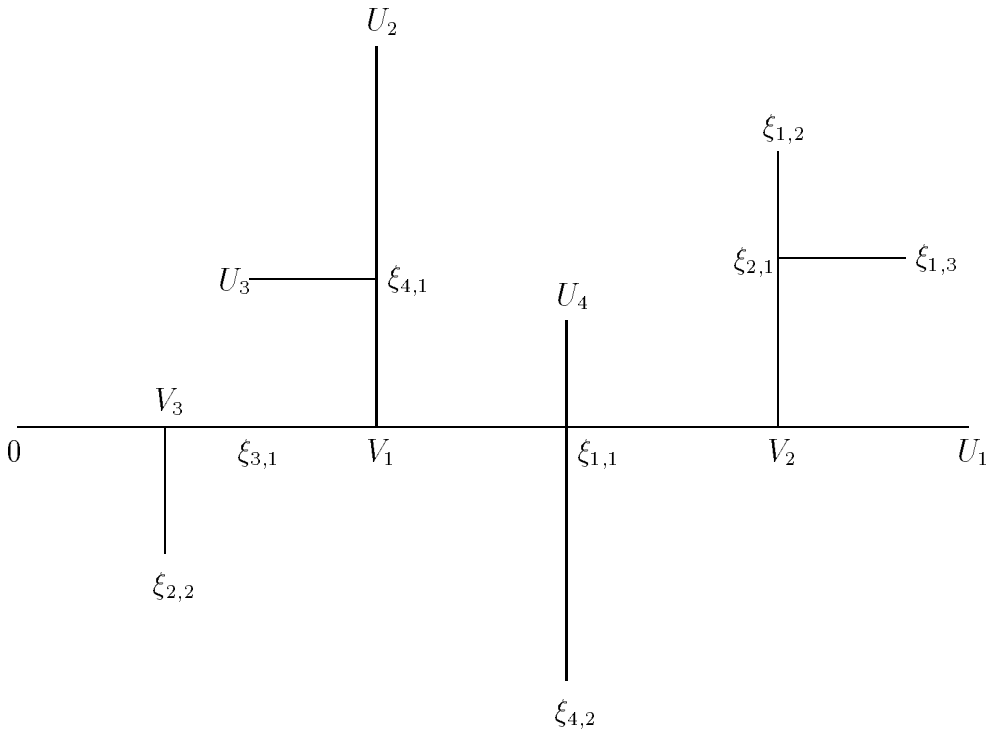

Figure 3

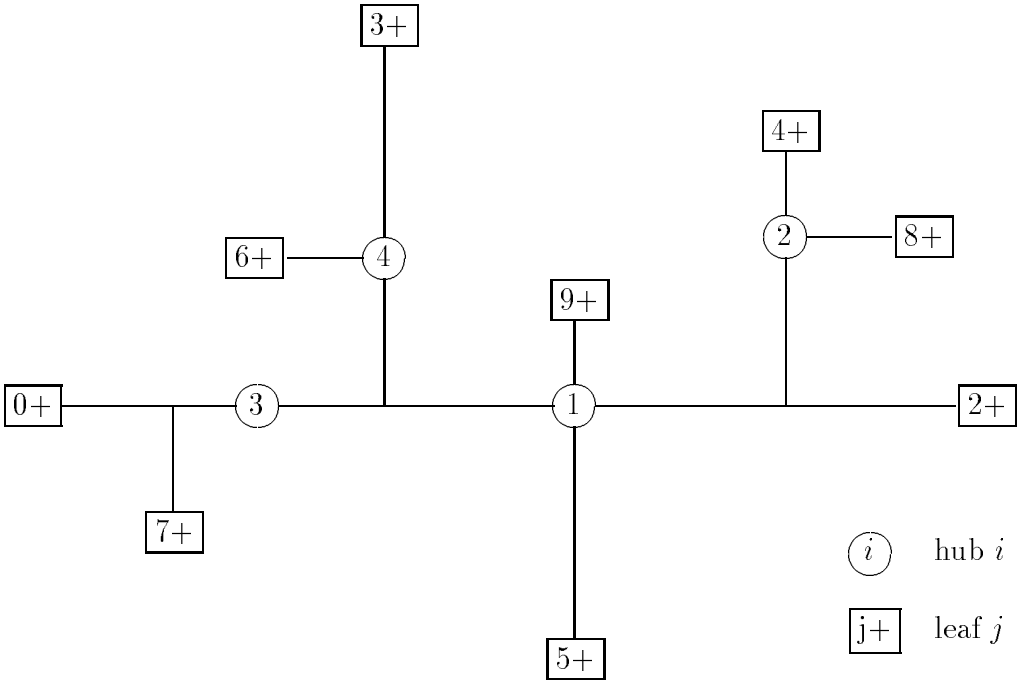

Figure 4

7.3 Consistency of the ICRT and the basic formula

Proposition 14 shows that a rescaling of the line-breaking construction gives a random tree with edge lengths distributed according to the basic formula. The proof uses a result from Camarri and Pitman [6] which exhibits the partial trees in the line-breaking construction as limits of spanning subtrees of p-trees.

Proposition 14 *Let* $\mathbf{c} = (c_1, \ldots, c_I)$, *where* $0 \leq I < \infty$ *and* $c_i > 0$ *for each* $1 \leq i \leq I$. *Set* $\theta := (1 + \sum_i c_i^2)^{-1/2}$. *For any* $k \geq 0$ *let* $\widetilde{\mathcal{S}}_{\theta \mathbf{c}[+k]}$ *denote the random tree constructed as in the previous section from parameters* $\theta_i := \theta c_i$, *and let* $\mathcal{S}_{\mathbf{c}[+k]}$ *be the tree whose distribution is specified by the basic formula for the vector* $\mathbf{c}[+k]$ *at* (30). *Then*

$$\mathcal{S}_{\mathbf{c}[+k]} \stackrel{d}{=} \theta \otimes \widetilde{\mathcal{S}}_{\theta \mathbf{c}[+k]}$$

where $\otimes$ *is the edge-scaling map from section 5.*

Equivalently, the continuum random tree $\widehat{\mathcal{T}}^{\mathbf{c}}$ whose construction was outlined in section 7.1 does indeed exist and can be represented as $\widehat{\mathcal{T}}^{\mathbf{c}} = \theta \otimes \mathcal{T}^{\theta \mathbf{c}}$ where $\mathcal{T}^{\theta \mathbf{c}}$ is the is ICRT $\mathcal{T}^{\boldsymbol{\theta}}$ of [2] for the vector $\boldsymbol{\theta} := \theta \mathbf{c}$, with $\theta_i := 0$ for $i > I$.

Proof. Given $\mathbf{c}[+k] \in \mathcal{C}$, define $p^{[n]}$ as at (7) in terms of $\mathbf{c}[+k]$. Write $\sigma_n = \sqrt{\sum_i (p_i^{[n]})^2}$. So

$$\sigma_n^2 = \sum_{i=1}^{I} \tfrac{c_i^2}{n} + o(\tfrac{1}{n}) + \tfrac{1 - n^{-1/2} c}{n - I} \sim \tfrac{1}{n}(1 + \sum_{i=1}^{I} c_i^2) = \tfrac{1}{n\theta^2}$$

and

$$\frac{p_i^{[n]}}{\sigma_n} \to \theta c_i, \ 1 \leq i \leq I; \ \ \frac{p_i^{[n]}}{\sigma_n} \to 0, I + 1 \leq i \leq I + k.$$

Recall from section 3.2 that $\mathcal{U}_n$ is the random $p^{[n]}$-tree and $\mathbf{s}(\mathcal{U}_n)$ is the subtree of $\mathcal{U}_n$ spanned by $[I + k]$. According to [6, Corollary 15]

$$\sigma_n \otimes \mathbf{s}(\mathcal{U}_n) \stackrel{d}{\to} \widetilde{\mathcal{S}}_{\theta \mathbf{c}[+k]}.$$

But Proposition 6 implies

$$n^{-1/2} \otimes \mathbf{s}(\mathcal{U}_n) \stackrel{d}{\to} \mathcal{S}_{\mathbf{c}[+k]}.$$

Since $n^{-1/2} \sim \theta \sigma_n$, the Proposition follows. □

Alternative Proof of a special case of Corollary 3 Suppose $c_j = c_k = 0$. By relabeling, we can assume $j = 1, k = 2$. Given $\mathbf{c}$, let $\theta := (1 + \sum_i c_i^2)^{-1/2}$ as in Proposition

14, and consider the line-breaking construction of $\widetilde{\mathcal{S}}_{\theta \mathbf{c}[+2]}$. The distance $\widetilde{L}_{12}$ between leaves 1+ and 2+ in $\widetilde{\mathcal{S}}_{\theta \mathbf{c}[+2]}$ is just the position η_1 of the first cutpoint in the construction. So by construction

$$P(\eta_1 > x) = \exp(-\tfrac{1}{2} a x^2) \prod_i (1 + \theta_i x) e^{-\theta_i x}$$

where $a = 1 - \sum_i \theta_i^2$. By Proposition 14, $L_{12} \stackrel{d}{=} \theta \widetilde{L}_{12}$. So

$$P(L_{12} > s) = P(\eta_1 > s/\theta) = \exp(-\frac{a s^2}{2\theta^2}) \prod_i (1 + c_i s) e^{-c_i s}. \tag{31}$$

But

$$\frac{a}{\theta^2} = \frac{1 - \sum_i (\theta c_i)^2}{\theta^2} = \frac{1}{\theta^2} - \sum_i c_i^2 = 1$$

so (31) is consistent with (4). □

Remark. As shown in [2], the line-breaking construction of the ICRT $\mathcal{T}^{\boldsymbol{\theta}}$ works not only for finite $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_I)$ but also for infinite $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ with $\sum_i \theta_i^2 \le 1$. While the combinatorial methods of this paper do not apply directly to the infinite $\boldsymbol{\theta}$, results in the infinite case can typically be deduced by approximation arguments with finite $\boldsymbol{\theta}$. For instance, there are an analogs of formulae (4) and (28) in the infinite case with finite sums and products replaced by infinite sums and products.

7.4 Distributional aspects of eternal additive coalescents

Fix $\mathbf{c} \in \mathcal{C}$ and $0 < \lambda < \infty$. For each $k \ge 2$, create a Poisson (rate λ per unit length) process of "cuts" along the edges of $\mathcal{S}_{\mathbf{c}[+k]}$. This creates a forest, and we can write

$$\mathbf{Y}^{\mathbf{c},k}(\lambda) = (Y_i^{\mathbf{c},k}(\lambda), i \ge 1)$$

for the vector of proportions of the k leaves $\{1+, \ldots, k+\}$ in the different tree-components, where the vector is written in decreasing order. It is shown in [2] that as $k \to \infty$ there is a limit random vector $\mathbf{Y}^{\mathbf{c}}(\lambda)$, which can also be obtained by a construction involving cutting along the skeleton of the continuum random tree $\widehat{\mathcal{T}}^{\mathbf{c}}$. The process $(\mathbf{Y}^{\mathbf{c}}(\lambda), 0 < \lambda < \infty)$ arises in [2] as the solution to a certain problem ("find all extreme versions of the additive coalescent"), but this solution is not very explicit, and it would be desirable to understand the distribution of $\mathbf{Y}^{\mathbf{c}}(\lambda)$ for given $\mathbf{c}$ and λ. In the special

case $\mathbf{c} = \mathbf{0}$ a description is given in [4], but the general case seems intractible. Some partial information about the distribution can be obtained as follows. For $k \geq 2$ write

$$M_{\mathbf{c}}^{(k)}(\lambda) = E \sum_i (Y_i^{\mathbf{c}}(\lambda))^k.$$

Then by (29) we can reinterpret $M_{\mathbf{c}}^{(k)}(\lambda)$ as the probability of the event that $\{1+, \ldots, k+\}$ are all in the same component of $\mathcal{S}_{\mathbf{c}[+k]}$. This event occurs if and only if there are no cuts within the spanning tree of $\{1+, \ldots, k+\}$, and so

$$M_{\mathbf{c}}^{(k)}(\lambda) = E \exp(-\lambda L(\mathcal{S}_{\mathbf{c}[+k]}^{[+k]}))$$

where $L(\mathcal{S}_{\mathbf{c}[+k]}^{[+k]})$ is the length of the spanning tree of $\mathcal{S}_{\mathbf{c}[+k]}$ spanned by $\{1+, \ldots, k+\}$. This provides motivation for the study of $L(\mathcal{S}_{\mathbf{c}[+k]}^{[+k]})$, mentioned in section 6.

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* Dover, New York, 1965.

[2] D. Aldous and J. Pitman. Inhomogeneous continuum random trees and the entrance boundary of the additive coalescent. Technical Report 525, Dept. Statistics, U.C. Berkeley, 1998. Available via http://www.stat.berkeley.edu/users/pitman.

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

[4] 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.

[5] L. Alonso and R. Schott. *Random Generation of Trees.* Kluwer, 1995.

[6] M. Camarri and J. Pitman. Limit distributions and random trees derived from the birthday problem with unequal probabilities. Technical Report 523, Dept. Statistics, U.C. Berkeley, 1998. Available via http://www.stat.berkeley.edu/users/pitman.

[7] 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).

[8] R. Durrett, H. Kesten, and E. Waymire. On weighted heights of random trees. *J. Theoretical Probab.*, 4:223–237, 1991.

[9] E.N. Gilbert and H.O. Pollak. Steiner minimal trees. *SIAM J. Appl. Math.*, 16:1–29, 1968.

[10] F.K. Hwang, D.S. Richards, and P. Winter. *The Steiner tree problem*, volume 53 of *Annals of Discrete Mathematics.* North Holland, Amsterdam, 1992.

[11] D.E. Knuth. *The Art of Computer Programming*, volume 1. Addison-Wesley, 1968.

[12] W.P. Maddison and M. Slatkin. Null models for the number of evolutionary steps in a character on a phylogenic tree. *Evolution*, 45:1184–1197, 1991.

[13] H. M. Mahmoud. *Evolution of Random Search Trees.* Wiley, 1992.

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

[15] J. Pitman. Exchangeable and partially exchangeable random partitions. *Probab. Th. Rel. Fields*, 102:145–158, 1995.

[16] 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.

[17] 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.

[18] J. Pitman. Partition structures derived from Brownian motion and stable subordinators. *Bernoulli*, 3:79–96, 1997.

[19] 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. To appear in *Combinatorics, Probability and Computing*.

[20] 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.

[21] 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.

[22] J.M. Steele. *Probability Theory and Combinatorial Optimization*. Number 69 in CBMS-NSF Regional Conference Series in Applied Math. SIAM, 1997.

[23] S. Tavaré. Line-of-descent and genealogical processes and their applications in population genetics. *Theoret. Population Biol.*, 26:119–164, 1984.