Electronic Journal of Probability

Vol. 5 (2000) Paper no. 2, pages 1–18.

Journal URL
http://www.math.washington.edu/~ejpecp/
Paper URL
http://www.math.washington.edu/~ejpecp/EjpVol5/paper2.abs.html

LIMIT DISTRIBUTIONS AND RANDOM TREES DERIVED FROM THE BIRTHDAY PROBLEM WITH UNEQUAL PROBABILITIES

Michael Camarri and Jim Pitman
Department of Statistics, University of California
367 Evans Hall # 3860, Berkeley, CA 94720-3860
pitman@stat.berkeley.edu

Abstract Given an arbitrary distribution on a countable set S consider the number of independent samples required until the first repeated value is seen. Exact and asymptotic formulæ are derived for the distribution of this time and of the times until subsequent repeats. Asymptotic properties of the repeat times are derived by embedding in a Poisson process. In particular, necessary and sufficient conditions for convergence are given and the possible limits explicitly described. Under the same conditions the finite dimensional distributions of the repeat times converge to the arrival times of suitably modified Poisson processes, and random trees derived from the sequence of independent trials converge in distribution to an inhomogeneous continuum random tree.

Keywords Repeat times, point process, Poisson embedding, inhomogeneous continuum random tree, Rayleigh distribution

AMS subject classification 60G55, 05C05

Research supported in part by N.S.F. Grants DMS 92-24857, 94-04345, 92-24868 and 97-03691

Submitted to EJP on October 14, 1998. Final version accepted on October 18, 1999.

Contents

1 Introduction **2**

2 Overview of Results **3**

3 Combinatorial formulæ **6**

3.1 The exact distribution of R_m 6

3.2 Analysis of the tree 8

4 Limit distributions **9**

4.1 Poisson embedding 9

4.2 Asymptotics for R_1. 11

4.3 Asymptotics of Joint Distributions 12

4.4 Representations in the plane 14

5 Asymptotics for the tree **15**

1 Introduction

Recall the classical *birthday problem*: given that each day of the year is equally likely as a possible birthday, and that birthdays of different people are independent, how many people are needed in a group to have a better than even chance that at least two people have the same birthday? The well known answer is 23. Here we consider a number of extensions of this problem. We allow the "birthdays" to fall in some finite or countable set S and let their common distribution be arbitrary on this set. We generalize the birthday problem in this setting as follows: in a stream of people, what is the distribution of the number who arrive before the mth person whose birthday is the same as that of some previous person in the stream? Our main motivation for studying the distributions of these random variables, which we call *repeat times*, is that they arise naturally in the study of certain kinds of random trees.

The distribution of the first repeat time has been studied widely. By truncating the Taylor series of the generating function Gail et al [14] derived an approximate distribution and applied their result to a problem of cell culture contamination. Using Newton polynomials Stein [28] derived the same approximation and supplied an error bound. Mase [21] used similar techniques to derive an approximation (with bound) in connection with the number of surnames in Japan. See also [18].

In the *quota problem* each possible value j, is assigned a quota, say v_j, and the problem is to describe the distribution of the time that a quota is first met. If $v_j = 2$ for all j this is the time of the first repeat. Using the technique of embedding in a Poisson process, Holst [15, 16] found expressions for the moments of a general quota fulfilment time, and specialised to find the asymptotic distribution of the first k-fold repeat time with the assumption that the probabilities are uniform across values. Here we use a Poisson embedding to derive asymptotic repeat time distributions for an arbitrary sequence of underlying value distributions. These results can easily be extended to the setting of the general quota problem. Aldous [1] gave a heuristic derivation of the limiting distributions for k-fold repeats. The results of section 4 are extended to k-fold repeats in the companion paper [9].

The birthday problem can also be approached by counting the number of matched pairs in a set. Theorem 5.G in Barbour, Holst, Jansen [6] gives a Poisson approximation (with error bound) to the number of matched pairs, from which the "if" part of Corollary 5 below may be deduced.

2 Overview of Results

This section presents some of the main results of the paper, with pointers to following sections for details and further developments.

Let p be a probability distribution on a finite or countable set S with $p_s > 0$ for all $s \in S$. We refer to elements of S as *values*. Let $Y_0, Y_1, \ldots$ be i.i.d.(p), meaning independent and identically distributed with common distribution p. Let R_m be the time of the mth *repeat* in this sequence. That is R_m is the mth index n such that $Y_n \in \{Y_0, \ldots, Y_{n-1}\}$. Let

$$\mathcal{A}_m := \{Y_0, \ldots, Y_{R_m - 1}\} = \{Y_0, \ldots, Y_{R_m}\}$$

denote the random set of observed values at the time of the mth repeat.

For an arbitrary $A \subseteq S$ let $|A|$ denote its cardinality, and define

$$p_A := \sum_{i \in A} p_i \quad \text{and} \quad \Pi_A := \prod_{i \in A} p_i.$$

Section 3 derives some exact formulæ for the distribution of R_m by conditioning on $\mathcal{A}_m$. In particular, for the first repeat R_1 there are the formulæ

$$P[R_1 = k] = \sum_{|A|=k} k!\, \Pi_A p_A \tag{1}$$

$$P[R_1 \geq k] = \sum_{|A|=k} k!\, \Pi_A \tag{2}$$

where the sums are over all subsets A of S of size k. The Ath term in (1) is $P(\mathcal{A}_1 = A)$. For the second repeat

$$P[R_2 = k] = \sum_{|A|=k-1} \frac{k!}{2} \Pi_A p_A^2 \tag{3}$$

where the Ath term is $P(\mathcal{A}_2 = A)$. These formulæ allow random variables with the same distribution as R_m to be recognized in other contexts, where results of this paper concerning the asymptotic distribution of R_m may be applied.

In particular, the distribution of R_m arises in the study of random trees. Given a sequence of S-valued random variables $(Y_0, Y_1, \ldots)$ define a directed graph

$$\mathcal{T}(Y_0, Y_1, \ldots) := \{(Y_{j-1}, Y_j) : Y_j \notin \{Y_0, \ldots, Y_{j-1}\}, j \geq 1\}. \tag{4}$$

Then $\mathcal{T}(Y_0, Y_1, \ldots)$ is a random tree labelled by $\{Y_0, Y_1, \ldots\}$ with root Y_0. Intuitively, the tree grows along the sequence until it encounters a repeat, at which point it backtracks to the first occurrence of the repeated value and continues its growth from there. The random tree $\mathcal{T}(Y_0, Y_1, \ldots)$ has been studied for $(Y_0, Y_1, \ldots)$ a finite state Markov chain [8],[20, §6.1]. By specializing a general Markov chain formula to the present setting, and evaluating a constant of normalization by use of Cayley's multinomial expansion [26, 25], there is the following result, an alternative proof of which is indicated after Lemma 7.

Lemma 1 [13]*If $Y_0, Y_1, \ldots$ are i.i.d.(p) for p with finite support $S := \{i : p_i > 0\}$, then the random tree $\mathcal{T} := \mathcal{T}(Y_0, Y_1, \ldots)$ has the following distribution on the set $\mathbf{T}(S)$ of all rooted trees labelled by S.*

$$P(\mathcal{T} = \mathbf{t}) = \prod_{s \in S} p_s^{C_s \mathbf{t}} \qquad (\mathbf{t} \in \mathbf{T}(S)) \tag{5}$$

where $C_s\mathbf{t}$ is the number of children (out-degree) of s in $\mathbf{t}$.

Properties of these random trees are linked to repeat times via the following two results, which are proved in Section 3.2.

Theorem 2 *If $Y_0, Y_1, \ldots$ are i.i.d.(p) for an arbitrary discrete distribution p then $Y_{R_1-1}, Y_{R_2-1}, \ldots$ are i.i.d.(p) and this collection of random variables is independent of the random tree $\mathcal{T}(Y_0, Y_1, \ldots)$.*

For a discrete distribution p with support S call a random tree $\mathcal{T}$ labelled by S a *p-tree* if $\mathcal{T}$ has the same distribution as $\mathcal{T}(Y_0, Y_1, \ldots)$ for an i.i.d.(p) sequence (Y_i). For finite S, the distribution of a p-tree $\mathcal{T}$ on $\mathbf{T}(S)$ is given by formula (5). See [25, 23, 24] regarding p-trees and related models for random forests.

Corollary 3 *Suppose that $\mathcal{T}$ is a p-tree and that $Y_0, Y_1, Y_2, \ldots$ are i.i.d.(p) independent of $\mathcal{T}$. For $m = 1, 2, \ldots$ let $\mathcal{S}_m$ be the subtree of $\mathcal{T}$ spanned by the root of $\mathcal{T}$ and $Y_1, \ldots, Y_m$, and let $\mathcal{T}_m$ be the subtree of $\mathcal{T}(Y_0, Y_1, \ldots)$ with vertex set $\{Y_0, \ldots, Y_{R_m-1}\}$. Then there is the equality of joint distributions*

$$(Y_1, \ldots, Y_m; \mathcal{S}_m) \stackrel{d}{=} (Y_{R_1-1}, \ldots, Y_{R_m-1}; \mathcal{T}_m) \tag{6}$$

which also holds jointly as m varies. In particular, the number of vertices of $\mathcal{S}_m$ has the same distribution as the number $R_m - m + 1$ of vertices of $\mathcal{T}_m$, which is the number of distinct values before the mth repeat in an i.i.d.(p) sequence.

The joint distribution featured in (6) is described explicitly in Section 3.2 by formula (18). According to Corollary 3 for $m = 1$, the distribution of R_1 described by (1) and (2) is also the distribution of the number of vertices on the path from X_1 to X_2 in a p-tree, for X_1 and X_2 with distribution p picked independently of each other and of the tree. For p the uniform distribution on a finite set this is equivalent to the formula of Meir and Moon [22] for the distribution of the distance between two distinct points in a uniform random tree. Another random variable with the same distribution as R_1 is the number C of cyclic points generated by a random $M : S \to S$ such that the $M(s)$ are i.i.d.(p) as s ranges over S. Jaworski [17] obtained an equivalent of (1) with C in place of R_1 for finite S. As observed in [23], this identity in distribution is explained by Joyal's [19] bijection between S^S and $S \times S \times \mathbf{U}(S)$ where $\mathbf{U}(S)$ is the set of unrooted trees labelled by S.

Consider now the problem of describing the asymptotic distribution of the first repeat time R_1 in an i.i.d.(p) sequence, in a limiting regime with the probability distribution p depending on a parameter $n = 1, 2, \ldots$. By an appropriate relabeling of the set of possible values by positive integers, there is no loss of generality in supposing that the nth distribution is a *ranked discrete distribution* $(p_{ni}, i \geq 1)$, meaning that

$$p_{n1} \geq p_{n2} \geq \cdots \geq 0 \text{ and } \sum_{i=1}^{\infty} p_{ni} = 1.$$

For each n let $Y_{nj}, j \geq 0$ be i.i.d. with this distribution, and for $m \geq 1$ define R_{nm} to be the time of the mth repeat in the sequence $(Y_{nj}, j \geq 0)$. In the *uniform case*, when

$$p_{ni} = 1/n, \quad 1 \leq i \leq n, \tag{7}$$

it is elementary and well known [12, p. 83] that for all $r \geq 0$

$$\lim_{n\to\infty} P[R_{n1}/\sqrt{n} > r] = e^{-r^2/2}. \tag{8}$$

Consider more generally the problem of characterizing the set of all possible asymptotic distributions of R_{n1} derived from a sequence of ranked distributions $(p_{ni}, i \geq 1)$ with $p_{n1} \to 0$ as $n \to \infty$. A central result of this paper, established in Section 4.2, is the solution to this problem provided by the following theorem:

Theorem 4 *Let R_{n1} be the index of the first repeated value in an i.i.d. sequence with discrete distribution whose point probabilities in non-increasing order are $(p_{ni}, i \geq 1)$. Let*

$$s_n := \sqrt{\textstyle\sum_i p_{ni}^2} \quad \text{and} \quad \theta_{ni} := p_{ni}/s_n.$$

(i) *If*

$$p_{n1} \to 0 \text{ as } n \to \infty \text{ and } \theta_i := \lim\nolimits_n \theta_{ni} \text{ exists for each } i \tag{9}$$

then for each $r \geq 0$

$$\lim_{n\to\infty} P[s_n R_{n1} > r] = e^{-\frac{1}{2}(1-\sum_i \theta_i^2) r^2} \prod_i (1 + \theta_i r) e^{-\theta_i r}. \tag{10}$$

(ii) *Conversely, if there exist positive constants $c_n \to 0$ and d_n such that the distribution of $c_n(R_{n1} - d_n)$ has a non-degenerate weak limit as $n \to \infty$, then $p_{n1} \to 0$ and limits θ_i exist as in* (i)*, so the weak limit is just a rescaling of that described in* (i)*, with $c_n/s_n \to \alpha$ for some $0 < \alpha < \infty$, and $c_n d_n \to 0$.*

Thus for a general sequence of ranked discrete distributions $(p_{ni}, i \geq 1)$ with $p_{n1} \to 0$ the appropriate scaling constants for the first repeat times are $(s_n, n \geq 1)$. The quantity θ_{n1} measures the probability of the most probable value relative to this scaling. In particular, Theorem 4 shows when the limit distribution of R_{n1} is the same as in the uniform case:

Corollary 5 *With the notation of the previous theorem,*

$$\lim_{n\to\infty} P[s_n R_{n1} > r] = e^{-\frac{1}{2}r^2} \tag{11}$$

for all $r \geq 0$ if and only if both $p_{n1} \to 0$ and $\theta_{n1} \to 0$ as $n \to \infty$.

This limiting *Rayleigh distribution* is that of the first point of a Poisson process on $[0,\infty)$ of rate t at time t. It is implicit in the work of Aldous [3] that in the uniform case the rescaled repeat times $R_{n1}/\sqrt{n}, R_{n2}/\sqrt{n}, \ldots$ converge jointly in distribution to the arrival times of such a Poisson process. In Section 4.3 we establish a corresponding generalisation of Theorem 4:

Theorem 6 *In the asymptotic regime* (9)*, for each $m \geq 1$ there is the convergence of m-dimensional distributions*

$$(s_n R_{n1}, s_n R_{n2}, \ldots, s_n R_{nm}) \xrightarrow{\mathrm{d}} (\eta_1, \eta_2, \ldots, \eta_m)$$

where $0 < \eta_1 < \eta_2 < \cdots$ are the arrival times for the superposition of independent point processes $M^*, M_1^-, M_2^-, \ldots$ where M^* is a Poisson process on $[0,\infty)$ of rate $(1 - \sum_i \theta_i^2)t$ at time t and M_i^- is a homogeneous Poisson process on $[0,\infty)$ of rate θ_i, with its first point removed.

Theorem 14 in Section 4.4 presents a refinement of this result in terms of a family of point processes in the plane constructed from independent Poisson processes. A corollary of Theorem 14, presented in Section 5, describes a sense in which the sequence of random trees $\mathcal{T}(Y_{nj}, j \geq 0)$ converges in distribution in the same limit regime (9) to a continuum random tree (CRT) which can be constructed directly from the point processes in the plane. This leads to a new kind of CRT, an *inhomogeneous continuum random tree (ICRT)* $\mathcal{T}^{\boldsymbol{\theta}}$, parameterised by the ranked non-negative sequence $\boldsymbol{\theta} := (\theta_i, i \geq 1)$ with $\sum_i \theta_i^2 \leq 1$. See Aldous-Pitman [4] for the study of various distributional properties of the limiting ICRT $\mathcal{T}^{\boldsymbol{\theta}}$, and Aldous-Pitman [5] for the application of this ICRT to the study of a coalescent process.

3 Combinatorial formulæ

3.1 The exact distribution of R_m

Recall that $\mathcal{A}_m$ is the random set of observed values $\{Y_0, \ldots, Y_{R_m}\}$ up to the time R_m of the mth repeat in the sequence $(Y_0, Y_1, \ldots)$. Since $R_m = k$ if and only if $\{Y_0, \ldots, Y_{R_m}\}$ contains $k + 1 - m$ distinct values

$$P[R_m = k] = P[|\mathcal{A}_m| = k + 1 - m] = \sum_{|A|=k+1-m} P[\mathcal{A}_m = A]. \tag{12}$$

Thus to describe the distribution of R_m it is enough to describe the distribution of the random set $\mathcal{A}_m$.

If $\mathcal{A}_1 = A$ then the first $|A|$ values taken by the Y_i are distinct and exactly the values A. Note that $R_1 = |A|$. By independence,

$$P[R_1 = |A| \,|\, \{Y_0, \ldots, Y_{|A|-1}\} = A] = p_A$$

and hence

$$P[\mathcal{A}_1 = A] = |A|!\,\Pi_A p_A. \tag{13}$$

This yields formula (1). More generally, if $\mathcal{A}_m = A$ then $(Y_0, \dots, Y_{R_m - 1})$ contains each of the elements of A plus $m-1$ repeated values. Again Y_{R_m} takes a repeated value and so

$$P[\mathcal{A}_m = A] = P[\{Y_0, Y_1, \dots, Y_{|A|+m-2}\} = A] p_A.$$

In particular, $(Y_0, Y_1, \dots, Y_{R_2 - 1})$ contains exactly one repeated value. The number of permutations of k objects with two indistinguishable and the rest distinct is $k!/2!$, thus for an arbitrary set A

$$P[\mathcal{A}_2 = A] = \sum_{i \in A} \frac{(|A|+1)!}{2!} \Pi_A p_i p_A = \frac{(|A|+1)!}{2!} \Pi_A p_A^2. \tag{14}$$

Combined with (12) this yields (3).

Similarly, $(Y_0, Y_1, \dots, Y_{R_3 - 1})$ contains either one triple repeat or two values repeated once each. Hence

$$P[\mathcal{A}_3 = A] = \Pi_A p_A \left(\sum_{i \in A} \frac{(|A|+2)!}{3!} p_i^2 + \sum_{\{i,j\} \subseteq A} \frac{(|A|+2)!}{2!2!} p_i p_j \right) \tag{15}$$

which combines with (12) to give a formula for the distribution of R_3.

To present a general formula for the distribution of $\mathcal{A}_m$ we need some notation involving partitions. Let $\mathbf{a} := (a_1, a_2, \dots)$ be a non-increasing sequence of non-negative integers with $|\mathbf{a}| := a_1 + a_2 + \dots < \infty$ and $l(\mathbf{a}) := \max\{i : a_i \neq 0\}$. Call $\mathbf{a}$ a *partition of* $|\mathbf{a}|$ *into* $l(\mathbf{a})$ *parts.* Let $P_A^{\mathbf{a}}$ be the symmetric polynomial in $\{p_i : i \in A\}$ where in each term the coefficient is 1 and the indices are $a_1, a_2, \dots$. For example

$$P_A^{(1)} := p_A := \sum_{i \in A} p_i \quad \text{and} \quad P_A^{(2,1)} := \sum_{i \in A} \sum_{j \in A \setminus \{i\}} p_i^2 p_j.$$

By a straightforward extension of the argument which led for $m = 1, 2, 3$ to formulæ (13), (14) and (15) respectively, there is the following general formula: for $m \geq 1$

$$P[\mathcal{A}_m = A] = \Pi_A p_A \sum_{|\mathbf{a}| = m-1} \frac{(|A| + m - 1)!}{(a_1+1)!(a_2+1)! \cdots} P_A^{\mathbf{a}} \tag{16}$$

where the sum is over all partitions $\mathbf{a} = (a_1, a_2, \dots)$ of $m-1$. The distribution of R_m is now determined by summing over appropriate sets A, as in formula (12). Alternatively, an expression for the tail probabilities of R_m is obtained by conditioning on the partition of k induced by values of $Y_0, Y_1, \dots Y_{k-1}$. Thus for $m \geq 1$ and $k \geq 1$

$$P[R_m \geq k] = \sum_{\substack{|\mathbf{b}| = k, \\ l(\mathbf{b}) > k - m}} \frac{k!}{b_1! b_2! \cdots} P_S^{\mathbf{b}}. \tag{17}$$

where the sum is over all partitions $\mathbf{b} = (b_1, b_2, \dots)$ of k into more than $k - m$ parts. In the particular case $m = 1$ this gives formula (2).

3.2 Analysis of the tree

Recall the definition (4) of $\mathcal{T}(Y_0, Y_1, \ldots)$. Theorem 2 and Corollary 3 are obtained by letting $m \to \infty$ in the following Lemma. Define $\mathbf{T}^*(S)$ to be the set of all rooted trees labelled by some finite non-empty subset of S. For $\mathbf{t} \in \mathbf{T}^*(S)$ the set of *leaves* of $\mathbf{t}$ is the set of all vertices of $\mathbf{t}$ whose out-degree in $\mathbf{t}$ is zero.

Lemma 7 *Let $\mathcal{T}_m$ be the subtree of $\mathcal{T}(Y_0, Y_1, \ldots)$ whose set of vertices is $\{Y_0, Y_1, \ldots, Y_{R_m}\}$. Let $(y_i, 1 \leq i \leq m) \in S^m$. Then for each $\mathbf{t} \in \mathbf{T}^*(S)$ whose set of leaves is contained in the set $\{y_i, 1 \leq i \leq m\}$ and each y_{m+1} in the set $V(\mathbf{t})$ of vertices of $\mathbf{t}$,*

$$P(Y_{R_i-1} = y_i, 1 \leq i \leq m; Y_{R_m} = y_{m+1}; \mathcal{T}_m = \mathbf{t}) = \left(\prod_{i=1}^{m+1} p_{y_i}\right)\left(\prod_{v \in V(\mathbf{t})} p_v^{C_v \mathbf{t}}\right) \tag{18}$$

and

$$P(Y_{R_i-1} = y_i, 1 \leq i \leq m; \mathcal{T}_m = \mathbf{t}) = \left(\prod_{i=1}^{m} p_{y_i}\right)\left(\prod_{v \in V(\mathbf{t})} p_v^{C_v \mathbf{t}}\right) p_{V(\mathbf{t})}. \tag{19}$$

Proof. Observe first that $\mathcal{T}_m$ is identical to the subtree of $\mathcal{T}(Y_0, Y_1, \ldots)$ spanned by $\{Y_0, Y_{R_1-1}, \ldots, Y_{R_m-1}\}$. This is obvious for $m = 1$, and can be established by induction for $m = 2, 3, \ldots$. Suppose true for m. If $R_{m+1} = R_m + 1$ then both $Y_{R_{m+1}}$ and $Y_{R_m} = Y_{R_{m+1}-1}$ lie among the vertices of $\mathcal{T}_m$, so $\mathcal{T}_{m+1} = \mathcal{T}_m$ and the conclusion for $m+1$ instead of m is evident. If $R_{m+1} > R_m + 1$ then there is a stretch of novel values, followed by a repeat value $Y_{R_{m+1}}$. The set of vertices of $\mathcal{T}_{m+1}$ is therefore $\mathcal{T}_m \cup \{Y_{R_m+1}, \ldots, Y_{R_{m+1}-1}\}$ where $Y_{R_m+1}, \ldots, Y_{R_{m+1}-1}$ is the set of vertices along the unique path in $\mathcal{T}(Y_0, Y_1, \ldots)$ which connects $Y_{R_{m+1}-1}$ to $\mathcal{T}_m$. So $\mathcal{T}_{m+1}$ is spanned by $\mathcal{T}_m \cup \{Y_{R_{m+1}-1}\}$ and the desired result is again obtained for $m+1$ instead of m.

Essentially the same inductive argument shows that for each given sequence of values $(y_i, 1 \leq i \leq m) \in S^m$, each tree $\mathbf{t} \in \mathbf{T}^*(S)$ with a vertices whose set of leaves is contained in the set $\{y_i, 1 \leq i \leq m\}$, and each $y_{m+1} \in V(\mathbf{t})$, there is a unique sequence $(w_j, 0 \leq j \leq m+a-1)$ such that

$$(Y_{R_i-1} = y_i, 1 \leq i \leq m; Y_{R_m} = y_{m+1}; \mathcal{T}_m = \mathbf{t}) \Leftrightarrow (Y_j = w_j, 0 \leq j \leq m+a-1)$$

The probability of this event is therefore $\prod_{j=0}^{m+a-1} p_{w_j}$ and it is easily shown that this product can be rearranged as in the formula (18). Formula (19) now follows by summing (18) over all $v \in V(\mathbf{t})$. □

Proof of Lemma 1. Now S is finite. Let $\mathcal{T} := \mathcal{T}(Y_0, Y_1, \ldots)$ and observe that $\mathcal{T} = \mathcal{T}_m$ if $\{Y_{R_1-1}, \ldots, Y_{R_m-1}\} = S$. Fix $m \geq |S|$ and sum both sides of (19) over the set of all sequences $(y_i) \in S^m$ such that $\{y_1, \ldots, y_m\} = S$. The result is that for all trees $\mathbf{t} \in \mathbf{T}(S)$

$$P(\{Y_{R_1-1}, \ldots, Y_{R_m-1}\} = S, \mathcal{T} = \mathbf{t}) = P(\{Y_1, \ldots, Y_m\} = S) \prod_{v \in S} p_v^{C_v \mathbf{t}}.$$

Because it is assumed that $p_i > 0$ for all $i \in S$, as $m \to \infty$ each of the probabilities $P(\{Y_{R_1-1}, \ldots, Y_{R_m-1}\} = S)$ and $P(\{Y_1, \ldots, Y_m\} = S)$ converges to 1, and (5) follows. □

Proof of Theorem 2. For finite S this is obtained by a reprise of the previous argument, using formula (19) and Lemma 1. The result for infinite S follows using the fact that the σ-field generated by $\mathcal{T}_m$ increases to the σ-field generated by $\mathcal{T}(Y_0, Y_1, \ldots)$. □

Proof of Corollary 3. This follows immediately from Theorem 2 and the first sentence in the proof of Lemma 7. □

A check. Since Y_0 is the root of $\mathcal{T}(Y_0, Y_1, \ldots)$, it follows from Theorem 2 that Y_0 and Y_{R_1-1} are independent with distribution p. That is

$$P(Y_0 = y, Y_{R_1-1} = z) = p_y p_z \quad (y, z \in S). \tag{20}$$

This is obvious for $y = z$, because $(Y_0 = y, Y_{R_1-1} = y) = (Y_0 = y, Y_1 = y)$. Let $\mathcal{A}$ be the random set $\{Y_1, \ldots, Y_{R_1-2}\}$. Then it is easily seen that for $y \neq z$ and every finite subset A of $S - \{y, z\}$

$$P(Y_0 = y, Y_{R_1-1} = z, \mathcal{A} = A) = p_y p_z \, |A|! \, \Pi_A (p_A + p_y + p_z) \quad (y \neq z) \tag{21}$$

Now (20) for $y \neq z$ follows from (21) and the following formula, which is valid for every subset B of a countable set S, and every probability distribution p on S, with $\Pi_A := \prod_{i \in A} p_i$:

$$\sum_{A \subseteq S - B} |A|! \, \Pi_A (p_A + p_B) = 1 \tag{22}$$

where the sum is over all finite subsets A of $S - B$. To verify (22) it suffices to consider the case when B is a singleton, say $B = \{y\}$. Similarly to (13) for each finite subset A of $S - \{y\}$

$$P(Y_0 = y, \mathcal{A}_1 = A) = p_y |A|! \Pi_A (p_A + p_y)$$

and (22) for $B = \{y\}$ follows by summation over A.

4 Limit distributions

Throughout this section we work with the setting and notation introduced in Theorem 4.

4.1 Poisson embedding

Without loss of generality, it will be assumed from now on that the i.i.d. sequences $(Y_{nj}, j = 0, 1, \ldots)$ have been constructed as follows for all $n = 1, 2, \ldots$ by embedding in a Poisson process. Let N be a homogeneous Poisson process on $[0, \infty) \times [0, 1]$ of rate 1 per unit area, with points say $\{(S_0, U_0), (S_1, U_1), \ldots\}$ where $0 < S_0 < S_1 < \ldots$ are the points of a homogeneous Poisson process on $[0, \infty)$ of rate 1 per unit length, and the U_i are i.i.d. with uniform distribution on $[0, 1]$, independent of the S_i. Define

$$N(t) := N([0, t] \times [0, 1]) \quad \text{and} \quad N(t^-) := N([0, t) \times [0, 1]).$$

For $n \geq 1$ partition $[0, 1]$ into intervals $I_{n1}, I_{n2}, \ldots$ such that the length of I_{ni} is p_{ni}. For $n > 0$, $j \geq 0$ define

$$Y_{nj} = \sum_i i 1(U_j \in I_{ni}), \tag{23}$$

so for each n the $Y_{nj}, j \geq 0$ are i.i.d. with distribution $(p_{ni}, i \geq 1)$. Let $(R_{nm}, m \geq 1)$ mark the repeats in this sequence and let $(T_{nm}, m \geq 1)$ be the corresponding times within N, that is $T_{nm} := \inf\{t : N(t) > R_{nm}\}$ which implies

$$N(T_{nm}^-) = R_{nm}. \tag{24}$$

The next lemma allows us to deduce limits in distribution for the finite dimensional distributions of $(R_{nm}, m \geq 1)$ from corresponding limits in distribution of $(T_{nm}, m \geq 1)$.

Lemma 8 *If $p_{n1} \to 0$, then for each $m \geq 1$ there is the convergence in probability*

$$\frac{R_{nm}}{T_{nm}} \xrightarrow{P} 1 \quad \text{as } n \to \infty.$$

Proof. By the strong law of large numbers $N(t^-)/t$ converges almost surely to 1 as $t \to \infty$ and hence by (24) it suffices to show that T_{nm} converges in probability to infinity. Since $T_{n1} \leq T_{nm}$ for each $m \geq 1$ it is enough to consider $m = 1$. But formulæ (26) and (28) below imply that

$$|\log P(T_{n1} > t)| \leq \frac{t^2}{2} \frac{p_{n1}}{(1 - tp_{n1})} \text{ for } 0 \leq t < p_{n1}^{-1} \tag{25}$$

and the conclusion follows. □

To check (25), observe that since

$$T_{n1} = \inf\{t : \exists i \text{ with } N([0,t] \times I_{ni}) \geq 2\}$$

and for each n the restrictions of N to $[0, \infty) \times I_{ni}$ are independent Poisson processes for $i = 1, 2, \ldots$, for each $t \geq 0$

$$P(T_{n1} > t) = g(t; p_{n1}, p_{n2}, \ldots) := \prod_i (1 + p_{ni}t)e^{-p_{ni}t}. \tag{26}$$

More generally, for an arbitrary sequence of real numbers $\boldsymbol{\theta} := (\theta_1, \theta_2, \ldots)$ with $\sum_i \theta_i^2 < \infty$ and $t \geq 0$ we define

$$g(t; \boldsymbol{\theta}) := \prod_i (1 + \theta_i t)e^{-\theta_i t}.$$

The function $g(t; \boldsymbol{\theta})$ also arises in the theory of regularised determinants of Hilbert-Schmidt operators (Carleman [10], Simon [27]).

Lemma 9 *Let $\boldsymbol{\theta} := (\theta_1, \theta_2, \ldots)$ be such that $\theta_1 \geq \theta_2 \geq \cdots \geq 0$ and $\sum_i \theta_i^2 < \infty$. Then for $0 \leq t < \theta_1^{-1}$,*

$$\log g(t; \boldsymbol{\theta}) = -\frac{t^2}{2} \sum_i \theta_i^2 + \frac{t^3}{3} \sum_i \theta_i^3 - \cdots \tag{27}$$

where the series is absolutely convergent; consequently, for such t

$$|\log g(t; \boldsymbol{\theta})| \leq \frac{t^2}{2} \frac{\theta_1 \sum_i \theta_i}{(1 - t\theta_1)} \tag{28}$$

and

$$\left| \log g(t; \boldsymbol{\theta}) + \frac{t^2}{2} \sum_i \theta_i^2 \right| \leq \frac{t^3}{3} \frac{\theta_1 \sum_i \theta_i^2}{(1 - t\theta_1)}. \tag{29}$$

Proof. If $0 \le t\theta_1 < 1$ then also $0 \le t\theta_i < 1$ for all i, so the expansion $\log(1+z) = z - z^2/2 + z^3/3 - \cdots$ for $|z| < 1$ yields

$$\log g(t;\boldsymbol{\theta}) = \sum_i \left(-t\theta_i + t\theta_i - \frac{t^2}{2}\theta_i^2 + \frac{t^3}{3}\theta_i^3 - \cdots \right) \tag{30}$$

which becomes (27) after switching the order of summation. To justify the switch by absolute convergence, let $s^2 := \sum_i \theta_i^2$ and note that for $k \ge 2$

$$\sum_i \theta_i^k \le \theta_1^{k-2} \sum_i \theta_i^2 = \theta_1^{k-2} s^2.$$

Therefore

$$\sum_{k\ge 2} \sum_i \frac{(t\theta_i)^k}{k} \le \sum_{k \ge 2} \frac{\theta_1^{k-2} s^2 t^k}{2} = \frac{s^2 t^2}{2(1-\theta_1 t)} < \infty. \tag{31}$$

The estimates (28) and (29) follow easily by similar comparisons of (27) to a geometric series with common ratio $t\theta_1$. $\square$

4.2 Asymptotics for R_1.

Observe first that for

$$s_n := \sqrt{\textstyle\sum_i p_{ni}^2} \quad \text{and} \quad \theta_{ni} := p_{ni}/s_n$$

as in Theorem 4, formula (26) yields for $r \ge 0$

$$P(s_n T_{n1} > r) = g(t; \theta_{n1}, \theta_{n2}, \ldots) := \prod_i (1 + \theta_{ni} r) e^{-\theta_{ni} r} \tag{32}$$

As a simple special case of the following proof, the case of Theorem 4 (i) when $\theta_1 = 0$ and the conclusion is (11) follows immediately from this formula combined with the estimate (29) above and the substitution of T_{n1} for R_{n1} justified by Lemma 8.

Proof of Theorem 4 (i). Fix $r > 0$ and let j_r, n_r be such that $n > n_r$ implies $r\theta_{nj_r} < 1$. Clearly

$$\lim_{n\to\infty} \prod_{i \le j_r} (1 + \theta_{ni} r) e^{-\theta_{ni} r} = \prod_{i \le j_r} (1 + \theta_i r) e^{-\theta_i r}.$$

In view of (32) and Lemma 8 it only remains to show

$$\lim_{n\to\infty} \prod_{i > j_r} (1 + \theta_{ni} r) e^{-\theta_{ni} r} = e^{-\frac{1}{2}(1 - \sum_i \theta_i^2) r^2} \prod_{i > j_r} (1 + \theta_i r) e^{-\theta_i r}. \tag{33}$$

From the choice of j_r, if $n > n_r$ equation (27) implies

$$\log \left[\prod_{i > j_r} (1 + \theta_{ni} r) e^{-\theta_{ni} r} \right] = -\sum_{i > j_r} \theta_{ni}^2 \frac{r^2}{2} + \sum_{i > j_r} \theta_{ni}^3 \frac{r^3}{3} - \cdots$$

Similarly

$$\log\left[e^{-\frac{1}{2}(1-\sum_i \theta_i^2)r^2}\prod_{i>j_r}(1+\theta_i r)e^{-\theta_i r}\right] = -\left(1-\sum_{i\le j_r}\theta_i^2\right)\frac{r^2}{2}+\sum_{i>j_r}\theta_i^3\frac{r^3}{3}-\cdots$$

Now

$$\sum_{i>j_r}\theta_{ni}^2 = 1-\sum_{i\le j_r}\theta_{ni}^2 \to 1-\sum_{i\le j_r}\theta_i^2$$

and it is easily checked, using the bound $\theta_{ni}^m \le \theta_{ni}^2\theta_{nj}^{m-2}$ for $i\ge j$ with large j, and $\sum_i \theta_{ni}^2 = 1$ for all n, that for all $m>2$

$$\lim_{n\to\infty}\sum_{i>j_r}\theta_{ni}^m = \sum_{i>j_r}\theta_i^m.$$

The kind of bound used in equation (31) now allows the proof to be completed by dominated convergence □

Proof of Theorem 4 (ii). By consideration of subsequential limits and convergence of types [7, Theorem 14.2], it is easily seen that it suffices to establish the following lemma. □

Lemma 10 *If $\alpha>0$ and $\boldsymbol{\theta} := (\theta_i, i\ge 1)$ is a non-increasing sequence of reals with $\sum_i \theta_i^2 \le 1$ then $(\alpha,\boldsymbol{\theta})$ can be uniquely reconstructed from the function $r\mapsto h(\alpha r;\boldsymbol{\theta})$ for $r\in[0,\infty)$, where*

$$h(r;\boldsymbol{\theta}) := e^{-\frac{1}{2}(1-\sum_i \theta_i^2)r^2}\prod_i(1+\theta_i r)e^{-\theta_i r}. \tag{34}$$

Proof. From (27), for $0\le r\le(\alpha\theta_1)^{-1}$,

$$\log h(\alpha r;\boldsymbol{\theta}) = -\alpha^2 r^2/2+\alpha^3 r^3\sum_i\theta_i^3/3-\alpha^4 r^4\sum_i \theta_i^4/4+\cdots$$

So from the function $r\mapsto h(\alpha r;\boldsymbol{\theta})$ we can uniquely extract the sequence

$$\alpha,\ \textstyle\sum_i\theta_i^3, \sum_i\theta_i^4,\ldots$$

Let $(I_i, i\ge 0)$ be a partition of the unit interval such that the length of I_0 is $1-\sum_i\theta_i^2$ and the length of I_i is θ_i^2 for all $i\ge1$, and set $Z:=\sum_i\theta_i 1(U\in I_i)$, where U is a uniform$[0,1]$ random variable. Then $\sum_i\theta_i^{2+k} = E(Z^k)$ for $k=1,2,\ldots$. But these moments of Z uniquely determine the distribution of Z on $[0,1]$ and it is easily seen that this distribution uniquely determines the sequence $(\theta_1,\theta_2,\ldots)$ □

4.3 Asymptotics of Joint Distributions

We start by proving the particular case of Theorem 6 when $\theta_i=0$ for all $i\ge1$. That is:

Lemma 11 *Let M be an inhomogeneous Poisson process on $[0,\infty)$ of rate t at time t and let $\eta_1,\eta_2,\ldots$ be the arrival times of M. If $p_{n1}\to 0$ and $\theta_{n1}\to0$ as $n\to\infty$ then for each $m\ge1$, as $n\to\infty$*

$$(s_nR_{n1}, s_nR_{n2},\ldots,s_nR_{nm}) \xrightarrow{\mathrm{d}} (\eta_1,\eta_2,\ldots,\eta_m).$$

Proof. As in Section 4.1 let N be a homogeneous Poisson process on $[0,\infty)\times[0,1]$ of rate 1 per unit area. Let N_{ni} be N restricted to $[0,\infty)\times I_{ni}$, where I_{ni} is an interval of length p_{ni}. Let N_{ni}^{-} denote the process N_{ni} with its first point removed and let $N_{ni}^{-}(t) := N_{ni}^{-}([0,t])$. Consider counting processes $X_n := (X_n(t), t\geq 0)$ where

$$X_n(t) := \sum_i N_{ni}^{-}(t/s_n)$$

and the sum converges since it is bounded above by $N(t/s_n)$. The arrival times for X_n are $s_nT_{n1}, s_nT_{n2}, \ldots$ so by Lemma 8 and standard theory of weak convergence of point processes (Daley and Vere-Jones [11, Theorem 9.1.VI]) it is enough to show that the processes X_n converge weakly to M.

For $n,i\geq 1$ let $\mathcal{F}^{ni} := (\mathcal{F}_t^{ni}, t\geq 0)$ be the natural filtration of $N_{ni}(\cdot/s_n)$ and let $\mathcal{F}^n := (\mathcal{F}_t^n, t\geq 0)$ be the smallest filtration containing $\{\mathcal{F}^{ni} : i\geq 1\}$. Let $(C_{ni}(t), t\geq 0)$ be the compensator of $N_{ni}^{-}(\cdot/s_n)$ with respect to the filtration $\mathcal{F}^{ni}$ and $(C_n(t), t\geq 0)$ the compensator of X_n with respect to $\mathcal{F}^n$. Thus

$$C_n(t) = \sum_i C_{ni}(t). \tag{35}$$

The compensator of M with respect to its natural filtration is $C(t) := t^2/2$. By Theorem 13.4.IV of Daley and Vere-Jones [11] it is sufficient to show $C_n(t) \xrightarrow{P} t^2/2$ for $t>0$. Thus it is enough to show $EC_n(t)\to t^2/2$ and $\operatorname{Var} C_n(t)\to 0$ for $t>0$.

The process $N_{ni} := (N_{ni}(r), r\geq 0)$ is a homogeneous Poisson process of rate p_{ni}, with compensator $(p_{ni}r, r\geq 0)$. Thus $(N_{ni}(t/s_n), t\geq 0)$ has compensator $(\theta_{ni}t, t\geq 0)$. If T_{ni1} is the time of the first point of N_{ni} then $(N_{ni}^{-}(t/s_n), t\geq 0)$ counts only those points that arrive after $t = s_nT_{ni1}$. Hence

$$C_{ni}(t) = \theta_{ni}(t - s_nT_{ni1})^{+} \tag{36}$$

where s_nT_{ni1} has an exponential distribution with rate θ_{ni}. A little calculus and equations (35) and (36) yield

$$EC_n(t) = \sum_i (e^{-t\theta_{ni}} - 1 + t\theta_{ni}) \tag{37}$$

$$\operatorname{Var} C_n(t) = \sum_i (1 - e^{-2t\theta_{ni}} - 2t\theta_{ni}e^{-t\theta_{ni}}). \tag{38}$$

For $x\geq 0$ there are the elementary inequalities

$$(1-x/3)x^2/2 \leq e^{-x} - 1 + x \leq x^2/2 \tag{39}$$

and

$$1 - e^{-2x} - 2xe^{-x} \leq x^3/3 \tag{40}$$

which applied to (37) and (38) imply

$$(1-\theta_{n1}t/3)t^2/2 \leq EC_n(t) \leq t^2/2 \tag{41}$$

$$\operatorname{Var} C_n(t) \leq \sum_i \theta_{ni}^3 t^3/3 \leq \theta_{n1}t^3/3. \tag{42}$$

By hypothesis $\theta_{n1}\to 0$ as $n\to\infty$ and the proof is complete. □

Proof of Theorem 6. Let $(j_n, n \geq 1)$ be a sequence such that

$$\lim_{n\to\infty} \sum_{i \leq j_n} \theta_{ni}^2 = \sum_i \theta_i^2.$$

Define the process $X_n^* := (X_n^*(t), t \geq 0)$ to count only the repeats of value $j_n + 1$ and above in the sequence $Y_{n0}, Y_{n1}, \ldots$ and let $X_{ni} := (X_{ni}(t), t \geq 0)$ count the repeats of value i, that is

$$X_n^*(t) = \sum_{i > j_n} N_{ni}^-(t/s_n)$$

$$X_{ni}(t) = N_{ni}^-(t/s_n).$$

Clearly X_{ni} converges weakly to M_i^-. The natural scaling for X_n^* is not s_n but rather $s_n^* = \sqrt{\sum_{i>j_n} p_{ni}^2}$. If $\sum_i \theta_i^2 < 1$ a simple modification of Lemma 11 shows that the processes $(X_n^*(s_n t / s_n^*), t \geq 0)$ converge weakly to M, a Poisson process of rate t at t. By construction

$$\left(\frac{s_n^*}{s_n}\right)^2 \to 1 - \sum_i \theta_i^2$$

and hence X_n^* converges weakly to M^*. Independence then implies that

$$(X_n^*, X_{n1}, \ldots, X_{nj_n}, 0, 0, \ldots) \xrightarrow{\mathrm{d}} (M^*, M_1^-, M_2^-, \ldots)$$

as $n \to \infty$. The case $\sum_i \theta_i^2 = 1$ is simpler and left to the reader. $\square$

4.4 Representations in the plane

In this section we extend the result of Theorem 6 by considering the joint distributions of the repeat times and the corresponding first occurrence times. Let

$$\mathbb{G} := \{(x, y) : 0 \leq y \leq x\}.$$

The limiting process in Lemma 11 is the projection onto the first coordinate of a homogeneous process of rate 1 on the octant $\mathbb{G}$. We make this connection explicit as follows. For $n \geq 1, m \geq 1$ let J_{nm} be the first time at which the value repeated at R_{nm} occurred, that is

$$J_{nm} := \min\{j \geq 0 : Y_{nj} = Y_{nR_{nm}}\}.$$

Define $\mathbf{G}_n$to be the point process on $\mathbb{G}$ whose collection of points is

$$\mathbf{G}_n := \{(s_n R_{nm}, s_n J_{nm}), m \geq 1\}.$$

See Daley and Vere-Jones [11, Chapter 9] for a treatment of convergence concepts for point processes.

Lemma 12 *Let* $\mathbf{G}$ *be a Poisson process on the octant* $\mathbb{G}$ *whose intensity measure is Lebesgue measure. If* $p_{n1} \to 0$ *and* $\theta_{n1} \to 0$ *as* $n \to \infty$ *then* $\mathbf{G}_n$ *converges weakly to* $\mathbf{G}$.

Proof. Let K_{nm} be the time corresponding to J_{nm} in the Poisson embedding. It is easily seen using Lemma 8 that J_{nm} and K_{nm} have common asymptotics in any regime with $p_{n1} \to 0$. The claimed weak convergence therefore amounts to the following: for each $m \geq 1$

$$(s_n T_{n1}, s_n K_{n1}, \ldots, s_n T_{nm}, s_n K_{nm}) \xrightarrow{\mathrm{d}} (U_1, V_1, \ldots, U_m, V_m)$$

where $(U_m, V_m), m \geq 1$ are the points of $\mathbf{G}$ arranged so that the U_m are increasing. The distribution of $\mathbf{G}$ can be derived from the following two facts:
(a) $(U_m, m \geq 1)$ is the sequence of arrival times of a Poisson process of rate t at time t, and
(b) conditional on $(U_1, \ldots U_m)$ the random variables $V_1, \ldots V_m$ are independent and V_j is uniform on $[0, U_j]$ for each $j \geq 1$.
So it is enough to show that the limit of the $\mathbf{G}_n$ satisfies these conditions. Condition (a) follows from Lemma 11 and (b) can be seen as follows. Given that none of the first m repeats is a triple repeat, each of the pairs (T_{nj}, K_{nj}) is the first two points of some homogeneous Poisson process, so K_{nj} given T_{nj} is uniform on $[0, T_{nj}]$, and this feature passes easily to the limit. The argument is then completed by the following lemma. □

Lemma 13 [9] *Let $T_{n1}^{(3)}$ be the time of the first triple repeat, that is*

$$T_{n1}^{(3)} := \inf\{t : \exists i \text{ with } N([0,t] \times I_{ni}) \geq 3\}.$$

If $p_{n1} \to 0$ and $\theta_{n1} \to 0$ as $n \to \infty$ then for all $m \geq 1$,

$$\lim_{n\to\infty} P[T_{nm} \leq T_{n1}^{(3)}] = 1.$$

The result of Lemma 12 is extended to the setting of Theorem 4 as follows:

Theorem 14 *In the asymptotic regime* (9) *the point process $\mathbf{G}_n$ converges weakly to the superposition of independent point processes $M_0, M_1, M_2, \ldots$ where M_0 is a homogeneous Poisson process on $\mathbb{G}$ with intensity $1 - \sum_i \theta_i^2$ per unit area, and for $i \geq 1$ the set of points of M_i is $\{(\xi_{i,j}, \xi_{i,1}), j \geq 2\}$ where $0 < \xi_{i,1} < \xi_{i,2} < \ldots$ are the points of a homogeneous Poisson process on $(0, \infty)$ with intensity θ_i per unit length.*

Proof. This is a straightforward variation of the the proof of Theorem 6.

5 Asymptotics for the tree

Let $\mathcal{T}_{nm}$ denote the tree $\mathcal{T}_m$ derived as in Corollary 3 from an i.i.d. sequence $(Y_{nj}, j \geq 0)$ with distribution $(p_{ni}, i \geq 1)$. So $\mathcal{T}_{nm}$ is the subtree of $\mathcal{T}(Y_{nj}, j \geq 0)$ spanned by Y_{n0} and the $Y_{n,R_{ni}-1}$ for $1 \leq i \leq m$. Consider the behaviour of the trees $\mathcal{T}_{nm}$ in the asymptotic regime (9). In the uniform case (7), results of Aldous [2] describe the asymptotic behaviour of a suitably reduced version of $\mathcal{T}_{nm}$, with edge lengths normalized by $1/\sqrt{n}$, in terms of a *continuum random tree (CRT)*. It follows from the previous results that Aldous's description can be transferred to the case $p_{n1} \to 0$ and $\lim_n \theta_{n1} = 0$, with normalization of edge lengths of $\mathcal{T}_{nm}$ by $s_n := \sqrt{\sum_i p_{ni}^2}$ instead of $1/\sqrt{n}$, and with the same limiting CRT. We now describe the limiting behaviour of

$\mathcal{T}_{nm}$ in the more general case, with $p_{n1} \to 0$ and $\lim_n \theta_{ni} = \theta_i$ for all i. This leads to a new kind of CRT, an *inhomogeneous continuum random tree (ICRT)* $\mathcal{T}^{\boldsymbol{\theta}}$, parameterised by the ranked non-negative sequence $\boldsymbol{\theta} := (\theta_i, i \geq 1)$ with $\sum_i \theta_i^2 \leq 1$.

Following Aldous-Pitman [5], we first introduce an appropriate space of trees for the description of the limit process involved. For $k \geq 0$ and $m \geq 1$ let $\mathbf{T}_{k,m}$ be the space of trees such that

(i) there are exactly $m+1$ leaves (vertices of degree 1), labeled $0+, \ldots, m+$;

(ii) there may be extra labeled vertices, with distinct labels in $\{1, \ldots, k\}$;

(iii) there may be unlabeled vertices of degree 3 or more;

(iv) each edge e has a length l_e, where l_e is a strictly positive real number.

Let E_{nm} denote the event that the vertices Y_{n0} and $Y_{n,R_{ni}-1}$ for $1 \leq i \leq m$ are $m+1$ distinct *leaves* of $\mathcal{T}_{nm}$, where edge directions in $\mathcal{T}_{nm}$ are now ignored, so the root Y_{n0} of $\mathcal{T}_{nm}$ may be a leaf. It follows easily from the previous results that the event E_{nm} has probability approaching 1 in the limit. If E_{nm} occurs define a $\mathbf{T}_{k,m}$-valued random tree $\mathcal{R}_{nkm}$, as follows. First make $\mathcal{T}_{nm}$ into a "tree with edge-lengths" by assigning length $s_n := \sqrt{\sum_i p_{ni}^2}$ to each edge. Relabel vertex Y_{n0} as vertex $0+$ and, for each $1 \leq j \leq m$, relabel vertex Y_{n,R_j-1} as vertex $j+$. Of the remaining vertices, those with labels $1 \leq i \leq k$ retain the label, and the others are unlabeled. Finally, unlabeled vertices of degree 2 are deleted. More precisely, each maximal l-edge path joining such vertices is replaced by a single edge of length ls_n. The resulting tree is $\mathcal{R}_{nkm}$. See [5] for a more detailed account of this and the following construction, with diagrams. If E_{nm} does not occur, we set $\mathcal{R}_{nkm} = \partial$ for some conventional state ∂ not in $\mathbf{T}_{k,m}$. We call $\mathcal{R}_{nkm}$ a *reduced tree* derived from $\mathcal{T}_{nm}$. To discuss weak convergence of the distribution of $\mathcal{R}_{nkm}$ as $n \to \infty$, we put the following topology on $\mathbf{T}_{k,m}$, then add ∂ as an isolated point. Each tree $\mathbf{t} \in \mathbf{T}_{k,m}$ has a *shape* shape$(\mathbf{t})$, which is the combinatorial tree obtained by ignoring edge-lengths. The set $\mathbf{T}_{k,m}^{\text{shape}}$ of possible shapes is finite. One can formally regard $\mathbf{t}$ as a vector (shape$(\mathbf{t})$; l_e, e an edge of shape$(\mathbf{t})$) and thereby $\mathbf{T}_{k,m}$ inherits a topology from the discrete topology on $\mathbf{T}_{k,m}^{\text{shape}}$ and the usual product topology on $\mathbb{R}^d$.

By construction of $\mathcal{R}_{nkm}$, given that E_{nm} occurs, the total length of all edges of $\mathcal{R}_{nkm}$ is $s_n R_{nm}$. According to Theorem 6, in the limit regime (9) the distribution of this total length converges as $n \to \infty$ to the distribution of the time η_m of the mth arrival in a limiting point process. Theorem 14 allows this convergence in distribution of the total length of $\mathcal{R}_{nkm}$ to be strengthened to convergence in distribution of $\mathcal{R}_{nkm}$ to $\mathcal{R}_{km}^{\boldsymbol{\theta}}$ for a random element $\mathcal{R}_{km}^{\boldsymbol{\theta}}$ of $\mathbf{T}_{k,m}$ which can be constructed directly from the Poisson point processes featured in Theorem 14. We state this formally in Corollary 15 below, following the construction of $\mathcal{R}_{km}^{\boldsymbol{\theta}}$ in the next paragraph from the Poisson processes featured in Theorem 14.

Fix $\boldsymbol{\theta} := (\theta_1, \theta_2, \ldots)$ with $\theta_1 \geq \theta_2 \geq \cdots 0$ and $\sum_i \theta_i^2 \leq 1$, and define $a := 1 - \sum_i \theta_i^2$. So $0 \leq a \leq 1$. If $a > 0$ let $((U_j, V_j), 1 \leq j < \infty)$ be the points of the Poisson point process M_0 of rate a per unit area on the octant $\{(u,v) : 0 \leq v \leq u < \infty\}$, labeled so that $0 < U_1 < U_2 < \ldots$. In the case $a = 0$, ignore subsequent mentions of U_j and V_j. For each i such that $\theta_i > 0$, let $0 < \xi_{i,1} < \xi_{i,2} < \ldots$ be the points of the Poisson point process on $(0, \infty)$ of rate θ_i per unit length. 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 $j \geq 2$ (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 $i \geq 1$ the mean number of i-cutpoints in that interval equals $\theta_i x - (1-\exp(-\theta_i x)) \leq \theta_i^2 x^2$. We may therefore order the cutpoints as $0 < \eta_1 < \eta_2 < \ldots$, where $\eta_j \to \infty$ as $j \to \infty$. These η_j then have the same joint distribution as the η_j in Theorem 6. We now build a tree by starting with the branch $[0,\eta_1]$ and then, inductively on $j \geq 2$, attaching the left end of the branch $(\eta_{j-1},\eta_j]$ to the joinpoint η_{j-1}^* corresponding to the cutpoint η_{j-1}. After m steps of this process, the interval $[0,\eta_m]$ has been randomly cut up and reassembled to form a random tree $\mathcal{T}_m^{\boldsymbol{\theta}}$ say, with vertex set $[0,\eta_m]$, with $m+1$ leaves $0,\eta_1,\ldots,\eta_m$, and with a finite set of branchpoints $\{\eta_{j-1}^*, 2 \leq j \leq m\}$. For any finite subset F of $[0,\eta_m]$ such that F contains all the leaves and branchpoints of $\mathcal{T}_m^{\boldsymbol{\theta}}$, the tree $\mathcal{T}_m^{\boldsymbol{\theta}}$ can be regarded as a tree with edge-lengths and vertex set F. For each $k \geq 0$ and $m \geq 1$ let $\widehat{\mathcal{R}}_m^{\boldsymbol{\theta}}$ be the tree with edgelegths so obtained from $\mathcal{T}_m^{\boldsymbol{\theta}}$ and the almost surely finite set F_{km} defined as the union of the set of all leaves and branchpoints of $\mathcal{T}_m^{\boldsymbol{\theta}}$ and the set of all i-joinpoints $\xi_{i,1}$ with $\xi_{i,1} < \eta_m$ and $1 \leq i \leq k$. Finally, let $\mathcal{R}_{km}^{\boldsymbol{\theta}}$ be the random element of $\mathbf{T}_{k,m}$ derived from $\widehat{\mathcal{R}}_m^{\boldsymbol{\theta}}$ by relabeling F_{km} as follows: let the leaves $0,\eta_1,\ldots,\eta_m$ be relabeled by $0+,1+,\ldots,m+$; for each i with $\xi_{i,1} < \eta_m$ and $1 \leq i \leq k$ let the i-joinpoint $\xi_{i,1}$ be relabeled by i, and let all remaining elements of F_{km} (i.e. the 0-joinpoints and i-joinpoints with $i > k$ among $\{\eta_{j-1}^*, 2 \leq j \leq m\}$) be unlabeled. It can be checked that the various operations involved in this continuous analog of the construction of $\mathcal{R}_{nkm}$ are appropriately continuous except on a set of probability zero in the limiting construction. Thus from the joint convergence of point processes underlying Theorem 14 we obtain:

Corollary 15 *For each $k \geq 0$ and $m \geq 1$, in the asymptotic regime* (9)

$$\mathcal{R}_{nkm} \xrightarrow{\mathrm{d}} \mathcal{R}_{km}^{\boldsymbol{\theta}} \text{ on } \mathbf{T}_{k,m}.$$

The random trees $\mathcal{T}_m^{\boldsymbol{\theta}}$, just used in the construction of $\mathcal{R}_{km}^{\boldsymbol{\theta}}$, are subtrees of an infinite tree with vertex set $[0,\infty)$, which defines the ICRT $\mathcal{T}^{\boldsymbol{\theta}}$ of [5] by completion in the metric on $[0,\infty)$ defined by path lengths in the infinite tree. The reduced trees $\mathcal{R}_{km}^{\boldsymbol{\theta}}$ then describe a consistent collection of finite-dimensional features of the infinite-dimensional ICRT $\mathcal{T}^{\boldsymbol{\theta}}$. See [5, 4] for further developments.

References

[1] D. Aldous. *Probability approximations via the Poisson clumping heuristic*, volume 77 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1989.

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

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

[4] D. Aldous and J. Pitman. A family of random trees with random edge lengths. *Random Structures and Algorithms*, 15:176–195, 1999.

[5] 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. To appear in *Prob. Th. and Rel. Fields.* Available via `http://www.stat.berkeley.edu/users/pitman`.

[6] A. D. Barbour, L. Holst, and S. Janson. *Poisson Approximation.* Clarendon Press, 1992.

[7] P. Billingsley. *Probability and Measure.* Wiley, New York, 1995. 3rd ed.

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

[9] M. Camarri. Asymptotics for k-fold repeats in the birthday problem with unequal probabilities. Technical Report 524, Dept. Statistics, U.C. Berkeley, 1998. Available via `http://www.stat.berkeley.edu`.

[10] T. Carleman. Zur Theorie der Linearen Integralgleichungen. *Math. Zeit*, 9:196–217, 1921.

[11] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes.* Springer-Verlag, Berlin, 1988.

[12] R. Durrett. *Probability: Theory and Examples.* Wadsworth-Brooks/Cole, 1995. 2nd ed.

[13] S. Evans and J. Pitman. Construction of Markovian coalescents. *Ann. Inst. Henri Poincaré*, 34:339–383, 1998.

[14] M. H. Gail, G. H. Weiss, N. Mantel, and S. J. O'Brien. A solution to the generalized birthday problem with application to allozyme screening for cell culture contamination. *J. Appl. Probab.*, 16(2):242–251, 1979.

[15] L. Holst. On birthday, collectors', occupancy and other classical urn problems. *International Statistical Review*, 54:15 – 27, 1986.

[16] L. Holst. The general birthday problem. *Random Structures Algorithms*, 6(2-3):201–208, 1995. Proceedings of the Sixth International Seminar on Random Graphs and Probabilistic Methods in Combinatorics and Computer Science, "Random Graphs '93" (Poznań, 1993).

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

[18] K. Joag-Dev and F. Proschan. Birthday problem with unlike probabilities. *Amer. Math. Monthly*, 99:10 – 12, 1992.

[19] A. Joyal. Une théorie combinatoire des séries formelles. *Adv. in Math.*, 42:1–82, 1981.

[20] 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.

[21] S. Mase. Approximations to the birthday problem with unequal occurrence probabilities and their application to the surname problem in Japan. *Ann. Inst. Statist. Math.*, 44(3):479–499, 1992.

[22] A. Meir and J. Moon. The distance between points in random trees. *J. Comb. Theory*, 8:99–103, 1970.

[23] 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`.

[24] 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`.

[25] J. Pitman. Coalescent random forests. *J. Comb. Theory A.*, 85:165–193, 1999.

[26] 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.

[27] B. Simon. *Functional integration and quantum physics*, volume 86 of *Pure and applied mathematics.* Academic Press, New York, 1979.

[28] C. Stein. Application of Newton's identities to a generalized birthday problem and to the Poisson Binomial distribution. Technical Report 354, Dept. Statistics, Stanford University, 1990.