Uncovering a Random Tree

Benjamin Hackl
Uppsala University, Sweden
University of Klagenfurt, Austria

Alois Panholzer
TU Wien, Austria

Stephan Wagner
Uppsala University, Sweden

Abstract

We consider the process of uncovering the vertices of a random labeled tree according to their labels. First, a labeled tree with n vertices is generated uniformly at random. Thereafter, the vertices are uncovered one by one, in order of their labels. With each new vertex, all edges to previously uncovered vertices are uncovered as well. In this way, one obtains a growing sequence of forests. Three particular aspects of this process are studied in this extended abstract: first the number of edges, which we prove to converge to a stochastic process akin to a Brownian bridge after appropriate rescaling. Second, the connected component of a fixed vertex, for which different phases are identified and limiting distributions determined in each phase. Lastly, the largest connected component, for which we also observe a phase transition.

2012 ACM Subject Classification Mathematics of computing → Random graphs; Mathematics of computing → Generating functions

Keywords and phrases Labeled tree, uncover process, functional central limit theorem, limiting distribution, phase transition

Digital Object Identifier 10.4230/LIPIcs.AofA.2022.10

Funding *Stephan Wagner*: supported by the Knut and Alice Wallenberg Foundation.

1 Introduction

We consider the process of uncovering the vertices of a random tree: starting either from one of the n^{n-2} unrooted or one of the n^{n-1} rooted unordered labeled trees of size n (i.e., with n vertices) chosen uniformly at random, we uncover the vertices one by one in order of their labels. This yields a growing sequence of (rooted) forests induced by the uncovered vertices, and we are interested in the evolution of these forests from the first vertex to the point that all vertices are uncovered.

This model is motivated by stochastic models known as coalescent models for particle coalescence, most notably the additive and the multiplicative coalescent [2] and the Kingman coalescent [7]. To make the distinction between these classical coalescent models and our model more explicit, let us briefly revisit the additive coalescent model (see [1]) as a prototypical example. This model describes a Markov process on a state space consisting of tuples $(x_1, x_2, \dots)$ with $x_1 \geq x_2 \geq \cdots \geq 0$ and $\sum_{i\geq 0} x_i = 1$ that model the fragmentation of a unit mass into clusters of mass x_i. Pairs of clusters with masses x_i and x_j then merge into a new cluster of mass $x_i + x_j$ at rate $x_i + x_j$. In the corresponding discrete time version of the process, exactly two clusters are merged in every time step. There are various combinatorial settings in which this discrete additive coalescent model appears, for example in the evolution of parking blocks in parking schemes related to "hashing with linear probing" [4], as the dual fragmentation process in the context of the random cutting of trees [6], and in a certain scheme for merging forests by uncovering one edge in every time step [10].

While the latter incarnation of the additive coalescent in which edges are uncovered successively is very much related in spirit to our vertex uncover model, the underlying processes are fundamentally different: these classical coalescent models rely on the fact that exactly two clusters are merged in every time step, which is not the case in our model. When uncovering a new vertex, a more or less arbitrary number of edges (including none at all) can be uncovered. There are coalescent models like the Λ-coalescent, a generalization of the Kingman coalescent [11], which allow for more than two clusters being merged – however, at present we are not aware of any known coalescent model that is able to capture the behavior of the vertex uncover process.

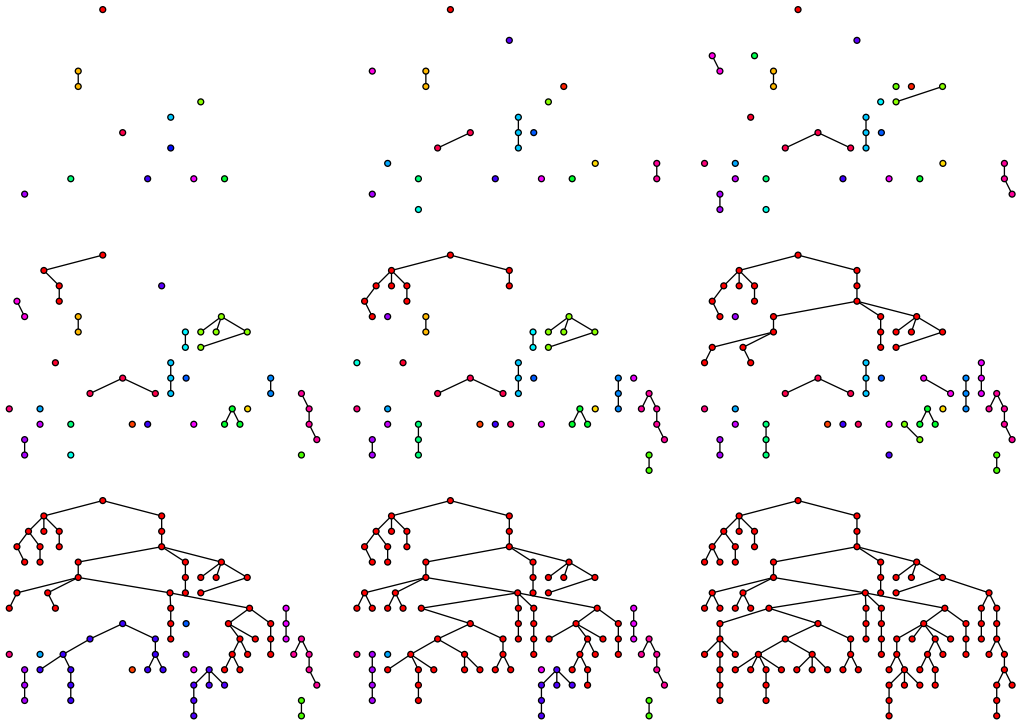

Figure 1 A few snapshots of the *uncover process* applied to a random labeled tree of size 100. From left to right and top to bottom, there are 12, 23, 34, ..., 89, and 100 uncovered vertices in the figures, respectively. Vertex labels are omitted for the sake of readability, and vertices are colored per connected component.

Overview. Different aspects of the uncover process on labeled trees are investigated in this extended abstract. In Section 2, we study the stochastic process given by the number of uncovered edges. The corresponding main result, a full characterization of the process and its limiting behavior, is given in Theorem 5. In this extended abstract, we sketch the proofs of the results in Section 2 – full details can be found in Appendix A.

Sections 3 and 4 are both concerned with cluster sizes, i.e., with the sizes of the connected components that are created throughout the process. In particular, in Section 3 we shift our attention to rooted labeled trees, to study the behavior of the component containing a fixed vertex. The expected size of the root cluster is analyzed in Theorem 9. Furthermore, we show that the number of rooted trees whose root cluster has a given size is given by a rather

simple enumeration formula – which, in turn, manifests in Theorem 11, a characterization of the different limiting distributions for the root cluster size depending on the number of uncovered vertices.

Finally, in Section 4 we use the results on the root cluster to draw conclusions regarding the size of the largest cluster in the tree.

Notation. Throughout this work we use the notation $[n] = \{1, \dots, n\}$ and $[k, \ell] = \{k, k+1, \dots, \ell\}$ for discrete intervals, and $x^{\underline{j}} = x(x-1)\cdots(x-j+1)$ for the falling factorials. The floor and ceiling function are denoted by $\lfloor x \rfloor$ and $\lceil x \rceil$, respectively. Furthermore, we use $\mathcal{T}$ and $\mathcal{T}^\bullet$ for the combinatorial classes of labeled trees and rooted labeled trees, respectively, and $\mathcal{T}_n$ and $\mathcal{T}_n^\bullet$ for the classes of labeled and rooted labeled trees of size n, i.e., with n vertices. Finally, we use $X_n \xrightarrow{d} X$ and $X_n \xrightarrow{p} X$ to denote convergence in distribution resp. probability of a sequence of random variables (r.v.) $(X_n)_{n\geq 0}$ to the r.v. X.

2 The number of uncovered edges

In this section our main interest is the behavior of the number of uncovered edges in the uncover process. We begin by introducing a formal parameter for this quantity.

▶ **Definition 1.** *Let T be a labeled tree with vertex set $V(T) = [n]$. For $1 \leq j \leq n$, we let $k_j(T) := \|T[1, 2, \dots, j]\|$ denote the number of edges in the subgraph of T induced by the vertices in $[j]$. We refer to the sequence $(k_j(T))_{1\leq j\leq n}$ as the* (edge) uncover sequence.

We start with a few simple observations. First, for any labeled tree of order n we have $k_1(T) = 0$, as well as $k_n(T) = n - 1$. Second, as any induced subgraph of a tree is a forest, and as forests have the elementary property that the number of edges together with the number of connected components gives the order of the forest, we find that $j - k_j(T)$ is the number of connected components after uncovering the first j vertices of T. Figure 2 illustrates the progression of the number of edges and the number of connected components for $1 \leq j \leq 1000$ in a randomly chosen labeled tree on 1000 vertices.

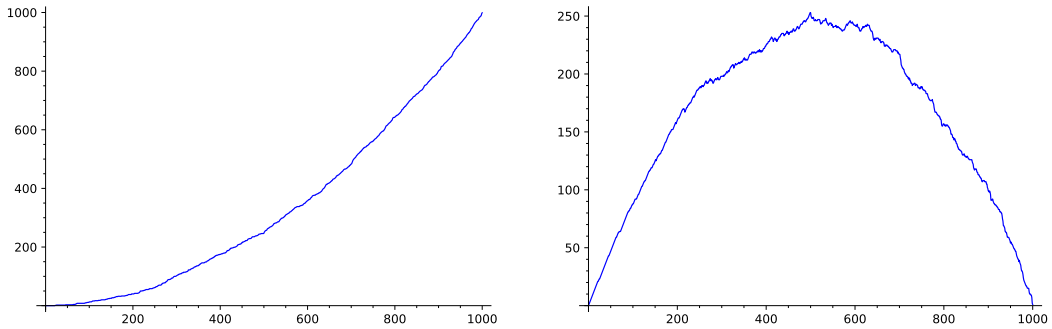

Figure 2 Progression of the number of edges (left) and the number of connected components (right) when sequentially uncovering a random labeled tree on 1000 vertices.

Moreover, the fact that the first j vertices of the tree induce a forest also yields the sharp bound $0 \leq k_j(T) \leq j - 1$ for all $1 \leq j \leq n - 1$. The lower bound is attained by the star with central vertex n, and the upper bound is attained by the (linearly ordered) path. We can also observe that as soon as $k_{n-1}(T) > 0$, the set of edges added in the last uncover step is not determined uniquely. Thus, the star with central vertex n is the only tree which is fully determined by its uncover sequence.

The following theorem provides explicit enumeration formulas for the number of trees with a partially and fully specified uncover sequence, respectively.

▶ **Theorem 2.** *Let r be a fixed positive integer with $1 \le r < n-1$, let $j_1, j_2, \ldots, j_r$ be positive integers with $1 < j_1 < j_2 < \cdots < j_r < n$, and let $a_1, a_2, \ldots, a_r$ be a nondecreasing sequence of nonnegative integers satisfying $a_i \le j_i - 1$ for all $1 \le i \le r$. Additionally, let $j_0 = 1$ and $a_0 = 0$. Then, the number of rooted labeled trees T of order n that satisfy $k_{j_i}(T) = a_i$ for all $1 \le i \le r$ is given by*

$$(n-j_r)^{j_r-a_r-1} n^{n-j_r-1} \times \prod_{i=1}^{r} \left(\sum_{h=0}^{a_i-a_{i-1}} \binom{j_{i-1}-a_{i-1}-1}{h} \binom{j_i-j_{i-1}}{a_i-a_{i-1}-h} (j_i-j_{i-1})^h j_i^{a_i-a_{i-1}-h} \right). \quad (1)$$

We first derive a helpful auxiliary result, namely an explicit formula for the corresponding (multivariate) generating function. The enumeration formula will then follow by extracting the appropriate coefficients.

▶ **Lemma 3.** *In the setting of Theorem 2, the multivariate generating function for the increments in the edge uncover process is given by*

$$E_n(z_1, z_2, \ldots, z_r) = n^{n-j_r-1} \prod_{i=1}^{r} \Big(n - j_r + j_i z_i + \sum_{h=i+1}^{r} (j_h - j_{h-1}) z_h \Big)^{j_i - j_{i-1}}. \quad (2)$$

In other words, the coefficient of the monomial $z_1^{a_1} z_2^{a_2-a_1} \ldots z_r^{a_r-a_{r-1}}$ in the expansion of $E_n(z_1, \ldots, z_r)$ is the number of labeled trees T of order n with $k_{j_i}(T) = a_i$ for all $1 \le i \le r$.

▶ Remark 4. By specifying the integers $j_1, j_2, \ldots, j_r$, the uncover process is effectively partitioned into intervals. This is also reflected by the quantities occurring in the product in (2): the difference $j_i - j_{i-1}$ corresponds to the number of vertices uncovered in the i-th interval, j_i represents the number of vertices uncovered in total up to the i-th interval, and $n - j_r$ corresponds to the number of vertices uncovered in the last interval.

Proof of Lemma 3. We begin by observing that when the process uncovers the vertex with label j, edges to all adjacent vertices whose label is less than j are uncovered as well. To determine the generating function of the edge increments, we assign the weight $x_i y_j$ to the edge connecting vertex i and vertex j with $i < j$, and then consider the generating function for the tree weight $w(T)$ (which is defined as the product of the edge weights); $E_n(z_1, \ldots, z_r) = \sum_{|T|=n} w(T)$.

Following Martin and Reiner [9, Theorem 4] or Remmel and Williamson [12, Equation (8)], the generating function of the tree weights $w(T)$ has the explicit formula

$$\sum_{|T|=n} w(T) = x_1 y_n \prod_{j=2}^{n-1} \Big(\sum_{i=1}^{n} x_{\min(i,j)} y_{\max(i,j)} \Big). \quad (3)$$

As initially observed, edges that are counted by $k_{j_i}(T)$ are precisely those that induce a factor y_ℓ for some $\ell \le j_i$. Thus we make the following substitutions: $x_\ell = 1$ for all ℓ, $y_\ell = z_i$ if and only if $j_{i-1} < \ell \le j_i$ (where[1] $j_0 = 1$), and $y_\ell = 1$ if $\ell > j_r$. To deal with the sum over $y_{\max(i,j)}$, observe that we can rewrite it as

$$\sum_{i=1}^{n} y_{\max(i,j)} = n - j_r + \sum_{i=1}^{j_1} y_{\max(i,j)} + \cdots + \sum_{i=j_{r-1}+1}^{j_r} y_{\max(i,j)}.$$

[1] Observe that y_1 does not occur, since at least one of the ends of every edge has a label greater than 1.

In this form, the different values assumed by the sum when j moves through the ranges $1 < j \leq j_1$, $j_1 < j \leq j_2$, etc. can be determined directly. For some j with $j_{i-1} < j \leq j_i$, the contribution to the product in (3) is

$$n - j_r + j_i z_i + \sum_{h=i+1}^{r} (j_h - j_{h-1}) z_h,$$

and for $j_r < j \leq n-1$ all y-variables are replaced by 1, so that the contribution to the product is a factor n. Putting both of these observations together shows that the right-hand side of (3) can be rewritten as the right-hand side of (2) and thus proves the lemma. ◀

With an explicit formula for the generating function of edge increments in the uncover process at our disposal, an explicit formula for the number of trees with given (partial) uncover sequence follows as a simple consequence.

Proof of Theorem 2. It remains to extract the coefficient of $z_1^{a_1} z_2^{a_2-a_1} \cdots z_r^{a_r-a_{r-1}}$, which is done step by step, starting with z_1. ◀

2.1 A closer look at the stochastic process

The exceptionally nice formula for the generating function of edge increments can be used to study the stochastic process that describes the number of uncovered edges in more detail. Let the sequence of random variables $(K_j^{(n)})_{1 \leq j \leq n}$ be the discrete stochastic process modeling the number of uncovered edges after uncovering the first j vertices in a random labeled tree of size n, chosen uniformly at random. The expected number of uncovered edges can be determined by a simple argument: with j uncovered vertices, $\binom{j}{2}$ of the $\binom{n}{2}$ possible positions for the edges have been uncovered. As every position is, due to symmetry and the uniform choice of the labeled tree, equally likely to hold one of the $n-1$ edges, we find

$$\mathbb{E}K_j^{(n)} = (n-1)\frac{\binom{j}{2}}{\binom{n}{2}} = \frac{j(j-1)}{n}. \tag{4}$$

To motivate our investigations further, consider the illustrations in Figure 3. The rescaled deviation from the mean is reminiscent of a stochastic process known as Brownian bridge.

In order to define this process formally, recall first that the Wiener process $(W(t))_{t \in [0,1]}$ is the unique stochastic process that satisfies

- $W(0) = 0$,
- W has independent, stationary increments,
- $W(t) \sim \mathcal{N}(0, t)$ for all $t > 0$, and
- $t \mapsto W(t)$ is almost surely continuous,

see [8, Definition 21.8]. A Brownian bridge can then be defined by setting

$$B(t) = (1-t)W(t/(1-t)), \tag{5}$$

see e.g. [13, Exercise 3.10]. The term "bridge" results from the fact that we have $B(0) = B(1) = 0$.

While the (normalized) deviation from the mean looks like it might converge to a Brownian bridge, we will prove that this is only *almost* the case. The following theorem characterizes the stochastic process. For technical purposes, we set $K_0^{(n)} = 0$ and introduce the linearly interpolated process $(\tilde{K}_t^{(n)})_{t \in [0,1]}$, where

$$\tilde{K}_t^{(n)} := (1 + \lfloor tn \rfloor - tn) K_{\lfloor tn \rfloor}^{(n)} + (tn - \lfloor tn \rfloor) K_{\lceil tn \rceil}^{(n)}, \tag{6}$$

which by construction has continuous paths.

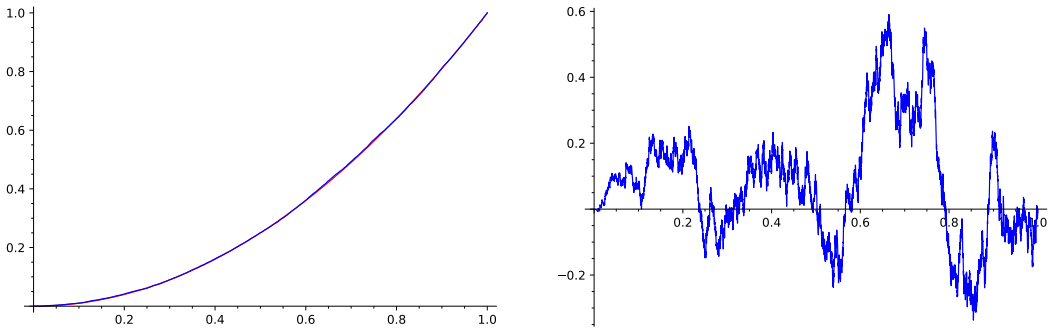

Figure 3 A path of the rescaled stochastic process $(K^{(n)}_{\lfloor tn \rfloor}/(n-1))_{t\in[0,1]}$ (left-hand side) and the corresponding (rescaled) deviation $\left(\frac{\tilde{K}^{(n)}_t - t^2 n}{\sqrt{n}}\right)_{t\in[0,1]}$ for a random labeled tree of size $n = 10000$.

▶ **Theorem 5.** *Let $(Z^{(n)}(t))_{t\in[0,1]}$ be the continuous stochastic process resulting from centering and rescaling the linearly interpolated process $(\tilde{K}^{(n)}_t)_{t\in[0,1]}$ in the form of*

$$Z^{(n)}(t) := \frac{\tilde{K}^{(n)}_t - t^2 n}{\sqrt{n}},$$

and let $(W(t))_{t\in[0,1]}$ be the standard Wiener process. Then, for $n \to \infty$, the rescaled process converges weakly with respect to the sup*-norm on $C([0,1])$ to a limiting process $Z^\infty(t)$ that is given by*

$$Z^\infty(t) = (1-t)W\big(t^2/(1-t)\big). \tag{7}$$

Furthermore, for s, $t \in [0,1]$ with $s < t$, the limiting process satisfies

$$\mathbb{E}Z^\infty(t) = 0, \quad \mathbb{V}Z^\infty(t) = t^2(1-t), \quad \text{and} \quad \operatorname{Cov}(Z^\infty(s), Z^\infty(t)) = s^2(1-t). \tag{8}$$

▶ Remark 6. While the limiting process $(Z^\infty(t))_{t\in[0,1]}$ is not a Brownian bridge (the corresponding variances and covariances as given in (8) do not match), it is closely related. Comparing the characterization of $Z^\infty(t)$ in (7) to (5), we see that the processes only differ by the square in the numerator of the argument of the Wiener process.

Two main ingredients are required to prove this result: a uniform tightness bound on the one hand, and information on the finite-dimensional joint distributions of $(\tilde{K}^{(n)}_t)_{t\in[0,1]}$ on the other hand.

To prove tightness, let us begin by revisiting (2). Given Cayley's well-known tree enumeration formula, the corresponding probability generating function for the complete uncover sequence, i.e., when we choose our integer vector as $\mathbf{j} = (2, 3, \ldots, n-1)$, is

$$P_n(z_2, \ldots, z_{n-1}) = \prod_{i=2}^{n-1} \Big(\frac{1}{n} + \frac{i}{n} z_i + \sum_{h=i+1}^{n-1} \frac{1}{n} z_h \Big). \tag{9}$$

This suggests modeling the process with $n-2$ independent random variables, each representing an edge increment[2]. The factorization suggests that the j-th increment (which corresponds to the factor with $i = j+1$) happens with probability $(j+1)/n$ when the vertex with label $j+1$

[2] We explicitly model edge increments here instead of edges, because with this approach we do not need to care about *which* edge is being uncovered. Our model explicitly only captures the behavior of the number of uncovered edges.

is uncovered, or with probability $1/n$ every time any of the subsequent vertices are uncovered. This probabilistic point of view can be used to construct a recursive characterization for the number of uncovered edges, namely[3]

$$K_{j+1}^{(n)} = K_j^{(n)} + \mathrm{Ber}\Big(\frac{j+1}{n}\Big) + \mathrm{Bin}\Big(j-1-K_j^{(n)}, \frac{1}{n-j}\Big). \tag{10}$$

The Bernoulli variable models the probability that the j-th edge increment is added when uncovering the vertex with label $j+1$, and the binomial variable models all of the remaining, not yet uncovered edge increments.

Now let us consider a centered and rescaled version of the process $(K_j^{(n)})_{1\le j\le n}$ by defining

$$Y_j^{(n)} := \frac{K_j^{(n)} - \frac{j(j-1)}{n}}{n-j}. \tag{11}$$

With the help of the recursive description in (10), one can show that $(Y_j^{(n)})_{1\le j\le n-1}$ is a martingale, see Appendix A.1.

We are now ready to state and prove the first required ingredient.

▶ **Lemma 7.** *For any real $C>1$ and any positive integer n, the random variable $Z^{(n)}(t)$ satisfies the tightness bound*

$$\mathbb{P}\big(\sup_{t\in[0,1]} |Z^{(n)}(t)| \ge C\big) \le 4(C-1)^{-2}, \tag{12}$$

so that for $C\to\infty$, the probability for the process to exceed C in absolute value converges to 0 uniformly in terms of n.

Sketch of proof. One first shows that $\sup_{t\in[0,1]}|Z^{(n)}(t)|$ can be bounded in terms of the deviation of the discrete process $(K_j^{(n)})_{0\le j\le n}$ from its mean. The bound then follows after expressing the discrete process in terms of the martingale $Y_j^{(n)}$ constructed above, partitioning the discrete interval $[0,n]$ appropriately and applying both Doob's L^p-inequality and the union bound. See Appendix A.2 for details. ◀

▶ **Lemma 8.** *Let r be a fixed positive integer, and let $\mathbf{t}=(t_1,\dots,t_r)\in(0,1)^r$. Then for $n\to\infty$, the random vector*

$$\mathbf{K}_{\lfloor \mathbf{t}n\rfloor}^{(n)} := (K_{\lfloor t_1 n\rfloor}^{(n)}, K_{\lfloor t_2 n\rfloor}^{(n)}, \dots, K_{\lfloor t_r n\rfloor}^{(n)})$$

converges, after centering and rescaling, for $n\to\infty$ in distribution to a multivariate normal distribution,

$$\frac{\mathbf{K}_{\lfloor \mathbf{t}n\rfloor}^{(n)} - \mathbb{E}\mathbf{K}_{\lfloor \mathbf{t}n\rfloor}^{(n)}}{\sqrt{n}} \xrightarrow[n\to\infty]{d} \mathcal{N}(\mathbf{0},\Sigma),$$

where the expectation vector $\mathbb{E}\mathbf{K}_{\lfloor \mathbf{t}n\rfloor}^{(n)}$ satisfies

$$\mathbb{E}\mathbf{K}_{\lfloor \mathbf{t}n\rfloor}^{(n)} = n(t_1^2, t_2^2, \dots, t_r^2) + O(1), \tag{13}$$

and the entries of the variance-covariance matrix $\Sigma=(\sigma_{i,j})_{1\le i,j\le r}$ are

$$\sigma_{i,j} = \begin{cases} t_i^2(1-t_j) & \text{if } i\le j,\\ t_j^2(1-t_i) & \text{if } i>j. \end{cases} \tag{14}$$

[3] We slightly abuse notation: formally, we would need to introduce auxiliary variables that are distributed according to the specified binomial and Bernoulli distributions.

Sketch of proof. The factorization of the probability generating function (9) implies that the distribution of the corresponding random vector $\Delta^{(n)}_{\lfloor \mathbf{t}n \rfloor} = (K^{(n)}_{\lfloor t_1 n \rfloor}, K^{(n)}_{\lfloor t_2 n \rfloor} - K^{(n)}_{\lfloor t_1 n \rfloor}, \dots, K^{(n)}_{\lfloor t_r n \rfloor} - K^{(n)}_{\lfloor t_{r-1} n \rfloor})$ is a marginal distribution of the sum of r independent, multinomially distributed random vectors. By the multivariate central limit theorem, $\Delta^{(n)}_{\lfloor \mathbf{t}n \rfloor}$ converges to a multivariate normal distribution – and as a consequence, so does $\mathbf{K}^{(n)}_{\lfloor \mathbf{t}n \rfloor}$.

The variance-covariance matrix of the centered and rescaled vector can be obtained, for example, by using the martingale constructed above. See again Appendix A.2 for details. ◀

Proof of Theorem 5. The proof relies on the well-known result asserting that given tightness of the sequence of corresponding probability measures as well as convergence of the finite-dimensional probability distributions, a sequence of stochastic processes converges to a limiting process (see [3, Theorem 7.1, Theorem 7.5]).

Tightness is implied by the uniform bound (12) derived in Lemma 7. The (limiting) behavior of the finite-dimensional distributions for the original process $(K^{(n)}_{\lfloor tn \rfloor})_{t\in[0,1]}$ is characterized by Lemma 8. This characterization carries over to the linearly interpolated process by an application of Slutsky's theorem [8, Theorem 13.18] after observing that

$$\mathbb{P}\left(\left|Z^{(n)}(t) - \frac{K^{(n)}_{\lfloor tn \rfloor} - t^2 n}{\sqrt{n}}\right| > \varepsilon\right) = \mathbb{P}\left(\frac{|\tilde{K}^{(n)}_t - K^{(n)}_{\lfloor tn \rfloor}|}{\sqrt{n}} > \varepsilon\right) \le \frac{\mathbb{E}((\tilde{K}^{(n)}_t - K^{(n)}_{\lfloor tn \rfloor})^2)}{n\varepsilon^2}$$
$$\le \frac{\mathbb{E}((K^{(n)}_{\lfloor tn \rfloor+1} - K^{(n)}_{\lfloor tn \rfloor})^2)}{n\varepsilon^2} \xrightarrow{n\to\infty} 0,$$

as a mechanical computation shows that $\mathbb{E}((K^{(n)}_{\lfloor tn \rfloor+1} - K^{(n)}_{\lfloor tn \rfloor})^2) = O(1)$.

Note that as the finite-dimensional distributions converge to Gaussian distributions, the limiting process $(Z^\infty(t))_{t\in[0,1]}$ is Gaussian itself – which means that it is fully characterized by its first and second order moments. As a consequence of Lemma 8, we find for all $s, t \in [0,1]$ with $s < t$ that

$$\mathbb{E}Z^\infty(t) = 0, \qquad \mathbb{V}Z^\infty(t) = t^2(1-t), \qquad \mathrm{Cov}(Z^\infty(s), Z^\infty(t)) = s^2(1-t).$$

It can be checked that if $(W(t))_{t\in[0,1]}$ is a standard Wiener process, the Gaussian process $((1-t)W(t^2/(1-t)))_{t\in[0,1]}$ has the same first and second order moments and therefore also the same distribution as Z^∞. ◀

3 Size of the root cluster

We now shift our attention from the number of uncovered edges to the sizes of the connected components (or *clusters*) appearing in the graph throughout the uncover process. It will prove convenient to change our tree model to *rooted* labeled trees, as the nature of rooted trees allows us to focus our investigation on one particular cluster – the one containing the root vertex. In case the root vertex has not yet been uncovered, we will consider the size of the root cluster to be 0. Formally, we let the random variable $R^{(k)}_n$ be the size of the root cluster of a (uniformly) random rooted labeled tree of size n with k uncovered vertices.

Using the symbolic method for labeled structures (cf. [5, Chapter II]), we can set up a formal specification for the corresponding combinatorial classes and subsequently extract functional equations for the associated generating functions. Let $\mathcal{T}^\bullet$ be the class of rooted labeled trees, and let $\mathcal{G}$ be a refinement of $\mathcal{T}^\bullet$ where the vertices can either be covered or uncovered, and where uncovered vertices are marked with a marker U. Finally, let $\mathcal{F}$ be

a further refinement of $\mathcal{G}$ where all uncovered vertices in the root cluster are additionally marked with marker V. A straightforward "top-down approach", i.e., a decomposition of the members of the tree family w.r.t. the root vertex, yields the formal specification

$$\mathcal{F} = \mathcal{Z} * \textsc{Set}(\mathcal{G}) + \mathcal{Z} \times \{U, V\} * \textsc{Set}(\mathcal{F}), \qquad \mathcal{G} = \mathcal{Z} * \textsc{Set}(\mathcal{G}) + \mathcal{Z} \times \{U\} * \textsc{Set}(\mathcal{G}).$$

Introducing the corresponding exponential generating functions $F := F(z, u, v)$ and $G := G(z, u)$, we obtain the characterizing equations

$$F = ze^G + zuve^F, \qquad G = z(1+u)e^G. \tag{15}$$

Of course, $G(z, u) = T^\bullet(z(1+u))$, where $T^\bullet$ is the exponential generating function associated with $\mathcal{T}^\bullet$, the Cayley tree function. Starting with (15), the following results on $R_n^{(k)}$ can be deduced.

▶ **Theorem 9.** *The expectation $\mathbb{E}(R_n^{(k)})$ is, for $0 \le k \le n$ and $n \ge 1$, given by*

$$\mathbb{E}(R_n^{(k)}) = \sum_{j=1}^{k} \frac{j\, k^{\underline{j}}}{n^j}. \tag{16}$$

Depending on the growth of $k = k(n)$, $\mathbb{E}(R_n^{(k)})$ has the following asymptotic behavior:

$$\mathbb{E}(R_n^{(k)}) \sim \begin{cases} \frac{k}{n}, & \text{for } k = o(n), \quad \textit{(k small)}, \\ \frac{\alpha}{(1-\alpha)^2}, & \text{for } k \sim \alpha n, \text{ with } 0 < \alpha < 1, \quad \textit{(k in central region)}, \\ \frac{n^2}{d^2}, & \text{for } k = n - d, \text{ with } d = \omega(\sqrt{n}) \text{ and } d = o(n), \\ & \quad \textit{(k subcritically large)}, \\ \kappa n, \quad \text{with} & \kappa = 1 - ce^{\frac{c^2}{2}} \int_c^\infty e^{-\frac{t^2}{2}} dt, \\ & \text{for } k = n - d, \text{ with } d \sim c\sqrt{n} \text{ and } c > 0, \quad \textit{(k critically large)}, \\ n - \sqrt{\frac{\pi}{2}} d\sqrt{n}, & \text{for } k = n - d, \text{ with } d = o(\sqrt{n}), \quad \textit{(k supercritically large)}. \end{cases}$$

Proof. After introducing $E := E(z, u) = \frac{\partial}{\partial v} F(z, u, v)\big|_{v=1} = \sum_{n,k} \frac{n^{n-1}}{n!} \binom{n}{k} \mathbb{E}(R_n^{(k)})$, considering the partial derivative of (15) with respect to v yields

$$E = \frac{1}{1 - \frac{u}{1+u} G} - 1.$$

Extracting coefficients and using $\mathbb{E}(R_n^{(k)}) = \frac{[z^n u^k] E}{[z^n u^k] G}$, we obtain (16). In order to analyze the asymptotic behavior of $\mathbb{E}(R_n^{(k)})$, the integral representation

$$\mathbb{E}(R_n^{(k)}) = \int_0^\infty (x-1)\, e^{-x} \Big(1 + \frac{x}{n}\Big)^k dx,$$

which can be verified in a straightforward way, turns out to be advantageous. Expanding the integrand and distinguishing several cases yields the asymptotic results given in the theorem. ◀

We can even obtain the exact distribution of $R_n^{(k)}$. There are two different approaches we want to briefly sketch: for one, an explicit formula for the generating function $F = F(z, u, v)$ can be found either from manipulating the recursive description (15), or directly by decomposing $\mathcal{F}$ as a tree forming the uncovered root cluster with a forest with covered roots attached. Either way, this yields

$$F = T^\bullet\big(vXe^{-X}\big) + \frac{G}{1+u}, \quad \text{with} \quad X = \frac{uG}{1+u}.$$

Extracting coefficients via an application of the Lagrange inversion formula (see, e.g., [5, Theorem A.2]) then yields an explicit formula for $F_{n,k,m} = n![z^n u^k v^m]F(z,u,v)$, i.e., the number of labeled rooted trees with n vertices of which k are uncovered and m belong to the root cluster (for $0 \le m \le k \le n$ and $n \ge 1$):

$$F_{n,k,m} = \begin{cases} \binom{n-1}{k} n^{n-1}, & m = 0, \\ \binom{n}{m}\binom{n-m-1}{k-m} n^{n-k-1} m^m (n-m)^{k-m}, & m \ge 1. \end{cases}$$

From this formula, the probabilities $\mathbb{P}(R_n^{(k)} = r)$ given below can be obtained directly.

Alternatively, there is also a more combinatorial approach to determine these probabilities: there is an elementary formula enumerating trees where a specified set of vertices forms a cluster.

▷ Claim 10. The number of trees on $[n]$ that do not have any edges between the vertex sets $[r]$ and $[r+1,k]$, where additionally the induced subgraph on $[r]$ is a tree itself, is given by

$$r^{r-1} n^{n-k-1} (n-k)(n-r)^{k-r-1}. \tag{17}$$

Proof of Claim 10. By Cayley's tree enumeration formula, the number of labeled trees on $[r]$ is r^{r-2}. Hence, the statement of the claim is equivalent to

$$r n^{n-k-1} (n-k)(n-r)^{k-r-1} \tag{18}$$

enumerating all rooted forests on $[n]$ whose roots are the vertices in $[r]$ that do not have any edges between $[r]$ and $[r+1,k]$. This can be proved bijectively with a construction similar to the Prüfer code, or by an application of Kirchhoff's Matrix-Tree Theorem [15, Theorem 5.6.8]. Details can be found in the full version of this extended abstract. ◁

As a consequence, the probability $\mathbb{P}(R_n^{(k)} = r)$ can be obtained by multiplying the number of these trees with $r\binom{k}{r}$ (to account for which vertices in $[k]$ form the root cluster and for the choice of the root), and then normalizing by n^{n-1}, the number of labeled rooted trees on n vertices.

▶ **Theorem 11.** *The exact distribution of $R_n^{(k)}$ is characterized by the following probability mass function (p.m.f.), which is given as follows for $0 \le m \le k \le n$ and $n \ge 1$ (and 0 otherwise):*

$$\mathbb{P}(R_n^{(k)} = m) = \begin{cases} 1 - \frac{k}{n}, & \text{for } m = 0, \\ \frac{m^m (n-k)(n-m)^{k-m-1}}{n^k}\binom{k}{m}, & \text{for } 1 \le m \le k < n, \\ 1, & \text{for } m = k = n. \end{cases}$$

Depending on the growth of $k = k(n)$, we obtain the following limiting behavior:

- *k small, i.e., $k = o(n)$:*

  $$R_n^{(k)} \xrightarrow{d} 0.$$

- *k in central region, i.e., $k \sim \alpha n$ with $0 < \alpha < 1$:*

  $$R_n^{(k)} \xrightarrow{d} R_\alpha, \quad \text{where the discrete r.v. } R_\alpha \text{ is characterized by its p.m.f.}$$

  $$\mathbb{P}(R_\alpha = m) =: p_m = \begin{cases} 1 - \alpha, & m = 0, \\ \frac{m^m}{m!}(1-\alpha)\alpha^m e^{-\alpha m}, & m \ge 1, \end{cases}$$

  or alternatively by the probability generating function $p(v) = \sum_{m \ge 0} p_m v^m = \frac{1-\alpha}{1 - T^\bullet(v\alpha e^{-\alpha})}$.

- *k subcritically large, i.e., $k = n - d$ with $d = \omega(\sqrt{n})$ and $d = o(n)$:*

$$\left(\frac{d}{n}\right)^2 \cdot R_n^{(k)} \xrightarrow{d} \textsc{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right),$$

where $\textsc{Gamma}(\frac{1}{2}, \frac{1}{2})$ is a Gamma-distribution characterized by its density $f(x) = \frac{1}{\sqrt{2\pi x}}\, e^{-\frac{x}{2}}$, for $x > 0$.

- *k critically large, i.e., $k = n - d$ with $d \sim c\sqrt{n}$ and $c > 0$:*

$$\frac{1}{n} \cdot R_n^{(k)} \xrightarrow{d} R(c),$$

where the continuous r.v. $R(c)$ is characterized by its density $f_c(x) = \frac{1}{\sqrt{2\pi}} \frac{c}{\sqrt{x}(1-x)^{\frac{3}{2}}}\, e^{-\frac{c^2 x}{2(1-x)}}$, for $0 < x < 1$.

- *k supercritically large, i.e., $k = n - d$ with $d = \omega(1)$ and $d = o(\sqrt{n})$:*

$$\frac{1}{d^2} \cdot \left(n - R_n^{(k)}\right) \xrightarrow{d} D,$$

where the continuous r.v. D is characterized by its density $f(x) = \frac{1}{\sqrt{2\pi}\, x^{\frac{3}{2}}}\, e^{-\frac{1}{2x}}$, $x > 0$.

- *k supercritically large with fixed difference, i.e., $k = n - d$ with d fixed:*

$$n - d - R_n^{(k)} \xrightarrow{d} D(d),$$

where the discrete r.v. $D(d)$ is characterized by the p.m.f.

$$\mathbb{P}(D(d) = j) =: p_j = e^{-d} \cdot \frac{d(d+j)^{j-1}}{j!} \cdot e^{-j}, \quad j \geq 0,$$

or alternatively via the probability generating function $p(v) = \sum_{j \geq 0} p_j v^j = e^{d(T^\bullet(\frac{v}{e}) - 1)}$.

Proof. The probability mass function of $R_n^{(k)}$ follows from the considerations made before the statement of the theorem. Due to its explicit nature, the limiting distribution results stated in Theorem 11 can be obtained in a rather straightforward way by applying Stirling's formula for the factorials after distinguishing several cases. ◀

▶ Remark 12. Of course, for labeled trees, the distribution of $R_n^{(k)}$ matches with the distribution of the cluster size of a random vertex. Furthermore, by conditioning, one can easily transfer the results of Theorem 11 to results for the size $S_n^{(k)}$ of the cluster of the k-th uncovered vertex: $\mathbb{P}(S_n^{(k)} = m) = \mathbb{P}(R_n^{(k)} = m | R_n^{(k)} > 0) = \frac{n}{k} \cdot \mathbb{P}(R_n^{(k)} = m)$, for $m \geq 1$.

4 Size of the largest uncovered component

With knowledge about the behavior of the root cluster at our disposal, we return to non-rooted labeled trees and study the size of the largest cluster. To this aim, we introduce the random variable $X_{n,r}^{(k)}$ which models the number of components of size r after uncovering the vertices 1 to k in a uniformly random labeled tree of size n.

Formally, $X_{n,r}^{(k)} \colon \mathcal{T}_n \to \mathbb{Z}_{\geq 0}$. Note that we have, for all labeled trees $T \in \mathcal{T}_n$,

$$\sum_{r=1}^{n} r \cdot X_{n,r}^{(k)}(T) = k. \tag{19}$$

▶ **Theorem 13.** *Let $n, k, r \in \mathbb{Z}_{\geq 0}$ with $0 \leq r \leq k \leq n$. The expected number of connected components of size r after uncovering k vertices of a labeled tree of size n chosen uniformly at random is*

$$\mathbb{E}X_{n,r}^{(k)} = \binom{k}{r}\left(\frac{r}{n}\right)^{r-1}\left(1-\frac{k}{n}\right)\left(1-\frac{r}{n}\right)^{k-r-1}. \tag{20}$$

Sketch of proof. Observe that $X_{n,r}^{(k)}$ can be written as a sum of Bernoulli random variables

$$X_{n,r}^{(k)} = \sum_{\substack{S\subseteq[k]\\|S|=r}} X_{n,S}^{(k)},$$

with $X_{n,S}^{(k)}$ being 0 or 1 depending on whether or not the vertices in S form a cluster after k uncover steps. By symmetry and linearity of the expected value, we have

$$\mathbb{E}X_{n,r}^{(k)} = \sum_{\substack{S\subseteq[k]\\|S|=r}} \mathbb{E}X_{n,S}^{(k)} = \binom{k}{r}\mathbb{E}X_{n,[r]}^{(k)}.$$

A formula for the expected value on the right-hand side follows from Claim 10, and thus proves the theorem. ◀

In the spirit of the observation in (19), the formula in Claim 10 provides a combinatorial proof for the following summation identity.

▶ **Corollary 14.** *Let $n, k \in \mathbb{Z}_{\geq 0}$ with $0 \leq k \leq n$. Then, the identity*

$$\sum_{r=1}^{k}\binom{k}{r}r^r n^{n-k-1}(n-r)^{k-r-1}(n-k) = kn^{n-2} \tag{21}$$

holds.

Proof. The right-hand side enumerates the vertices in $[k]$ in all labeled trees on n vertices. The left-hand side does the same, with the summands enumerating the vertices in connected components of size r. ◀

▶ Remark 15. Observe that the identity in (21) can be rewritten as

$$\sum_{r=1}^{k}\binom{k}{r}r^r(n-r)^{k-r-1} = \frac{k}{n-k}n^{k-1},$$

which is a specialized form of Abel's Binomial Theorem – a classical, and well-known result; see, e.g., [14].

For a tree $T \in \mathcal{T}$, let $c_{\max}^{(k)}(T)$ denote the largest connected component of T after uncovering the first k vertices.

▶ **Theorem 16.** *Let $n \in \mathbb{Z}_{\geq 0}$, and let $T_n \in \mathcal{T}_n$ be a tree chosen uniformly at random. Then the behavior of the random variable $c_{\max}^{(k)}(T_n)$ as $n \to \infty$ can be described as follows:*

- *for $k = n - d$ with $d = \omega(\sqrt{n})$ (*subcritical case*), we have $c_{\max}^{(k)}(T_n)/n \xrightarrow{p} 0$.*
- *for $k = n - d$ with $d \sim c\sqrt{n}$ for a constant c (*critical case*), the rescaled random variable $c_{\max}^{(k)}(T_n)/n$ converges weakly to a (non-degenerate) continuous limiting distribution,*
- *for $k = n - d$ with $d = o(\sqrt{n})$ (*supercritical case*), we have $c_{\max}^{(k)}(T_n)/n \xrightarrow{p} 1$. With high probability, there is one "giant" component whose size is asymptotically equal to n.*

Sketch of proof. For the subcritical case, we use the expected root cluster size from Theorem 9. Since a cluster of size r contains the root with probability $\frac{r}{n}$, we have

$$\frac{n^2}{d^2} \sim \mathbb{E}R_n^{(n-d)} = \sum_{r=0}^{n-d} \mathbb{E}X_{n,r}^{(n-d)} \cdot r \cdot \frac{r}{n} \geq \sum_{r=m}^{n-d} \mathbb{E}X_{n,r}^{(n-d)} \frac{r^2}{n} \geq \frac{m^2}{n} \sum_{r=m}^{n-d} \mathbb{E}X_{n,r}^{(n-d)}$$
$$\geq \frac{m^2}{n} \mathbb{P}(c_{\max}^{(n-d)}(T_n) \geq m).$$

This implies that

$$\mathbb{P}(c_{\max}^{(n-d)}(T_n) \geq m) = O\Big(\frac{n^3}{d^2 m^2}\Big),$$

so if $m = \epsilon n$ for any fixed $\epsilon > 0$, we have $\mathbb{P}(c_{\max}^{(n-d)}(T_n) \geq m) \to 0$.

In the critical case, we first consider the situation that there is a cluster that contains more than half of the vertices. Clearly, such a cluster must be the largest cluster and the only cluster of its size. Therefore, if $r = \rho n$ with $\rho > \frac{1}{2}$, we have

$$\mathbb{P}(c_{\max}^{(k)}(T_n) = r) = \mathbb{E}X_{n,r}^{(k)} \sim \frac{1}{n} \frac{e^{-\frac{c^2}{2}\frac{\rho}{1-\rho}}}{\sqrt{2\pi}} \frac{c}{(\rho(1-\rho))^{3/2}},$$

using Theorem 13 and Stirling's approximation. Thus, for $\rho > \frac{1}{2}$,

$$\mathbb{P}(c_{\max}^{(k)}(T_n) \geq \rho n) \to \int_{\rho}^{1} \frac{e^{-\frac{c^2}{2}\frac{t}{1-t}}}{\sqrt{2\pi}} \frac{c}{(t(1-t))^{3/2}}\, dt.$$

For $\rho \leq \frac{1}{2}$, we can modify this argument with a generalized version of Theorem 13 for several clusters and the inclusion-exclusion principle to prove convergence of $c_{\max}^{(k)}(T_n)/n$ to a continuous random variable with support $[0, 1]$. Details are left to the full version.

Finally, in the supercritical case, we recall the corresponding case for the size of the root cluster from Theorem 9. Using Markov's inequality yields, for any $\epsilon > 0$,

$$\mathbb{P}(n - R_n^{(k)} \geq \epsilon n) \leq \frac{n - \mathbb{E}(R_n^{(k)})}{\epsilon n} \sim \frac{d\sqrt{n}}{\epsilon n} \xrightarrow{n\to\infty} 0.$$

Thus, the root cluster is the largest cluster of size $\sim n$ with high probability. Translating this from rooted to unrooted trees proves the theorem. ◀

References

1 David Aldous and Jim Pitman. The standard additive coalescent. *Ann. Probab.*, 26(4):1703–1726, 1998. doi:10.1214/aop/1022855879.

2 Jean Bertoin. *Random fragmentation and coagulation processes*, volume 102 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2006. doi:10.1017/CBO9780511617768.

3 Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, Inc., New York, second edition, 1999. doi:10.1002/9780470316962.

4 P. Chassaing and G. Louchard. Phase transition for parking blocks, Brownian excursion and coalescence. *Random Structures Algorithms*, 21(1):76–119, 2002. doi:10.1002/rsa.10039.

5 Philippe Flajolet and Robert Sedgewick. *Analytic combinatorics*. Cambridge University Press, Cambridge, 2009. doi:10.1017/CBO9780511801655.

6 Svante Janson. Random cutting and records in deterministic and random trees. *Random Structures Algorithms*, 29(2):139–179, 2006. doi:10.1002/rsa.20086.

7 J. F. C. Kingman. The coalescent. *Stochastic Process. Appl.*, 13(3):235–248, 1982. doi:10.1016/0304-4149(82)90011-4.

8 Achim Klenke. *Probability theory—a comprehensive course*. Universitext. Springer, Cham, 2020. Third edition. doi:10.1007/978-3-030-56402-5.

9 Jeremy L. Martin and Victor Reiner. Factorization of some weighted spanning tree enumerators. *J. Combin. Theory Ser. A*, 104(2):287–300, 2003. doi:10.1016/j.jcta.2003.08.003.

10 Jim Pitman. Coalescent random forests. *J. Combin. Theory Ser. A*, 85(2):165–193, 1999. doi:10.1006/jcta.1998.2919.

11 Jim Pitman. Coalescents with multiple collisions. *Ann. Probab.*, 27(4):1870–1902, 1999. doi:10.1214/aop/1022677552.

12 Jeffery B. Remmel and S. Gill Williamson. Spanning trees and function classes. *Electron. J. Combin.*, 9(1):Research Paper 34, 24, 2002. URL: http://www.combinatorics.org/Volume_9/Abstracts/v9i1r34.html.

13 Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, third edition, 1999. doi:10.1007/978-3-662-06400-9.

14 John Riordan. *Combinatorial identities*. John Wiley & Sons, Inc., New York, 1968.

15 Richard P. Stanley. *Enumerative combinatorics. Vol. 2*, volume 62 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999. doi:10.1017/CBO9780511609589.

A Additional details on the proof of Theorem 5

A.1 Computations related to the martingale

With the help of the recursive description in (10), we can show that $(Y_j^{(n)})_{1\le j\le n-1}$ is a martingale by computing

$$\begin{aligned}\mathbb{E}(Y_{j+1}^{(n)}|Y_j^{(n)}) &= \frac{\mathbb{E}(K_{j+1}^{(n)}|K_j^{(n)})}{n-j-1} - \frac{j(j+1)}{n(n-j-1)} = \frac{K_j^{(n)} + \frac{j+1}{n} + \frac{j-1-K_j^{(n)}}{n-j}}{n-j-1} - \frac{j(j+1)}{n(n-j-1)}\\ &= \frac{K_j^{(n)}}{n-j} - \frac{(j-1)j}{n(n-j)} = Y_j^{(n)}.\end{aligned}$$

We can also give an explicit expression for the variance of $Y_k^{(n)}$: recall that by (11), we have $\mathbb{V}Y_k^{(n)} = (n-k)^{-2}\mathbb{V}K_k^{(n)}$. Then, with (10) and the laws of total variance and total expectation we find the recurrence

$$\mathbb{V}K_{k+1}^{(n)} = \Big(1 - \frac{1}{n-k}\Big)^2 \mathbb{V}K_k^{(n)} + \frac{(n-k-1)(2n-k-1)k}{(n-k)n^2},$$

for $1 \le k < n-1$ and $\mathbb{V}K_1^{(n)} = 0$. This allows us to conclude that

$$\mathbb{V}K_k^{(n)} = \sum_{j=1}^{k-1} \Big(\frac{n-k}{n-j-1}\Big)^2 \frac{(n-j-1)(2n-j-1)j}{(n-j)n^2} = \frac{k(k-1)(n-k)}{n^2}, \tag{22}$$

where the sum can be evaluated with the help of partial fractions and telescoping.

A.2 Proofs of auxiliary results

Proof of Lemma 7. In order to obtain the tightness condition, we show first that it can be reduced to an inequality for the martingale from the previous section. To this end, let us write $tn = j + \eta$, with $j \in \mathbb{Z}$ and $\eta \in [0, 1)$. A simple calculation shows that

$$\begin{aligned} Z^{(n)}(t) &= \frac{\tilde{K}_t^{(n)} - t^2 n}{\sqrt{n}} \\ &= \frac{(1-\eta)K_j^{(n)} + \eta K_{j+1}^{(n)} - (j+\eta)^2/n}{\sqrt{n}} \\ &= \frac{(1-\eta)(K_j^{(n)} - j(j-1)/n) + \eta(K_{j+1}^{(n)} - j(j+1)/n) - (j+\eta^2)/n}{\sqrt{n}} \\ &= (1-\eta)\frac{K_j^{(n)} - j(j-1)/n}{\sqrt{n}} + \eta\frac{K_{j+1}^{(n)} - j(j+1)/n}{\sqrt{n}} - \frac{j+\eta^2}{n^{3/2}}. \end{aligned}$$

The final fraction is bounded by 1, since $j + \eta^2 \le j + \eta = tn \le n$. It follows that

$$\sup_{t\in[0,1]} |Z^{(n)}(t)| \le \sup_{0\le j\le n} \Big|\frac{K_j^{(n)} - j(j-1)/n}{\sqrt{n}}\Big| + 1,$$

so

$$\begin{aligned} \mathbb{P}\Big(\sup_{t\in[0,1]} |Z^{(n)}(t)| \ge C\Big) &\le \mathbb{P}\Big(\sup_{0\le j\le n} \Big|\frac{K_j^{(n)} - j(j-1)/n}{\sqrt{n}}\Big| \ge C-1\Big) \\ &= \mathbb{P}\Big(\sup_{1\le j\le n-1} \Big|\frac{Y_j^{(n)}(n-j)}{\sqrt{n}}\Big| \ge C-1\Big). \end{aligned} \tag{23}$$

Note here that we need not consider $j = 0$ and $j = n$ in the supremum, since $K_j^{(n)} - j(j-1)/n = 0$ in either case. Since $(Y_j^{(n)})_{1\le j\le n-1}$ is a martingale, we can use Doob's L^p-inequality [8, Theorem 11.2]. For any real $C > 0$ and any fixed integer k with $1 \le k \le n-1$, we have

$$\mathbb{P}\Big(\sup_{1\le j\le k} |Y_j^{(n)}| \ge C\Big) \le \frac{\mathbb{V}Y_k^{(n)}}{C^2} = \frac{k(k-1)}{C^2(n-k)n^2}.$$

With this, we have all required prerequisites to prove tightness of $Z^{(n)}(t)$. We partition the interval over which the supremum is taken in (23), apply the martingale inequality, and then obtain the desired result after summing over all these upper bounds. For every integer $i > 0$, let $I_i^{(n)} := [2^{-i}n, 2^{-i+1}n] \cap \mathbb{Z}$. We find

$$\begin{aligned} \mathbb{P}\Big(\sup_{n-j\in I_i^{(n)}} \Big|\frac{Y_j^{(n)}(n-j)}{\sqrt{n}}\Big| \ge C-1\Big) &\le \mathbb{P}\Big(\sup_{n-j\in I_i^{(n)}} |Y_j^{(n)} 2^{-i+1}\sqrt{n}| \ge C-1\Big) \\ &= \mathbb{P}\Big(\sup_{n-j\in I_i^{(n)}} |Y_j^{(n)}| \ge \frac{2^{i-1}(C-1)}{\sqrt{n}}\Big) \\ &\le \frac{n}{2^{2i-2}(C-1)^2}\mathbb{V}(Y_{n-\lceil 2^{-i}n\rceil}^{(n)}) \\ &\le \frac{n}{2^{2i-2}(C-1)^2}\cdot\frac{2^i}{n} = \frac{4}{2^i(C-1)^2}, \end{aligned}$$

where in the last inequality we bounded the variance as follows:

$$\mathbb{V}(Y^{(n)}_{n-\lceil 2^{-i}n\rceil}) = \frac{\mathbb{V}(K^{(n)}_{n-\lceil 2^{-i}n\rceil})}{\lceil 2^{-i}n\rceil^2} = \frac{(n-\lceil 2^{-i}n\rceil)(n-\lceil 2^{-i}n\rceil-1)\lceil 2^{-i}n\rceil}{n^2\lceil 2^{-i}n\rceil^2} \leq \frac{2^i}{n}.$$

Finally, the union bound together with the observation that $\sum_{i\geq 1}\frac{4}{2^i(C-1)^2} = 4(C-1)^{-2}$ yields the upper bound in (12) and therefore completes the proof. ◀

Proof of Lemma 8. As a consequence of Lemma 3 and Cayley's well-known enumeration formula for labeled trees of size n, we find that the probability generating function of the number of edge increments after $1 < j_1 < j_2 < \cdots < j_r < n$ steps, respectively, is given by

$$\begin{aligned} P_n(z_1, z_2, \ldots, z_r) &= \frac{E_n(z_1, z_2, \ldots, z_r)}{n^{n-2}} \\ &= \prod_{i=1}^{r}\Big(1 - \frac{j_r}{n} + \frac{j_i}{n}z_i + \sum_{h=i+1}^{r}\frac{j_h - j_{h-1}}{n}z_h\Big)^{j_i - j_{i-1}}, \end{aligned} \tag{24}$$

where $j_0 = 1$ for the sake of convenience. Now observe that $\Delta^{(n)}_{\mathbf{j}}$ can be seen as a marginal distribution of the sum of r independent, multinomially distributed random vectors: write $t_i = j_i/n$ and consider $M_j \sim \mathrm{Multi}(j_i - j_{i-1}, \mathbf{p}_i)$ where

$$\mathbf{p}_i = (p_{i,0}, p_{i,1}, \ldots, p_{i,r}) \in [0,1]^r \quad \text{such that} \quad p_{i,h} = \begin{cases} 1 - t_r & \text{if } h = 0, \\ 0 & \text{if } 0 < h < i, \\ t_i & \text{if } h = i, \\ t_h - t_{h-1} & \text{otherwise.} \end{cases} \tag{25}$$

By construction, the probability generating function of M_i is then given by

$$\Big((1-t_r)z_0 + t_i z_i + \sum_{h=i+1}^{r}(t_h - t_{h-1})\Big)^{j_i - j_{i-1}},$$

so that the probability generating function of the sum $M_1 + \cdots + M_r$ is a product that is very similar (and actually equal if we set $z_0 = 1$, which corresponds to marginalizing out the first component) to (24). In order to make the following arguments formally easier to read, and as the first component is not relevant for us at all, we slightly abuse notation and let M_i for $1 \leq i \leq r$ denote the corresponding marginalized multinomial distributions instead.

For the sake of convenience, we make a slight adjustment: instead of fixing the integer vector $\mathbf{j} = (j_1, \ldots, j_r)$, we fix $\mathbf{t} = (t_1, \ldots, t_r)$ with $0 < t_1 < \cdots < t_r < 1$ and define $\mathbf{j} = \lfloor \mathbf{t}n \rfloor$. Here, n is considered to be sufficiently large so that the conditions for the corresponding integer vector, $1 < \lfloor t_1 n\rfloor < \cdots < \lfloor t_r n\rfloor < n$, are still satisfied.

By the multivariate central limit theorem, it is well-known that a multinomially distributed random vector $M \sim \mathrm{Multi}(n, \mathbf{p})$ converges, for $n \to \infty$ and after appropriate scaling, in distribution to a multivariate normal distribution,

$$\frac{M - n\mathbf{p}}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathrm{diag}(\mathbf{p}) - \mathbf{p}^\top\mathbf{p}). \tag{26}$$

As a consequence, we find that

$$\begin{aligned}\frac{\Delta^{(n)}_{\lfloor n\mathbf{t}\rfloor}-\mathbb{E}\Delta^{(n)}_{\lfloor n\mathbf{t}\rfloor}}{\sqrt{n}} &= \frac{(M_1+\cdots+M_r)-\mathbb{E}(M_1+\cdots+M_r)}{\sqrt{n}}\\ &= (\sqrt{t_1}+O(n^{-1}))\frac{M_1-\mathbb{E}M_1}{\sqrt{\lfloor t_1 n\rfloor}}+\cdots\\ &\quad+(\sqrt{t_r-t_{r-1}}+O(n^{-1}))\frac{M_r-\mathbb{E}M_r}{\sqrt{\lfloor t_r n\rfloor-\lfloor t_{r-1}n\rfloor}}\\ &\xrightarrow[n\to\infty]{d}\sqrt{t_1}\mathcal{N}(\mathbf{0},\Sigma_1)+\cdots+\sqrt{t_r-t_{r-1}}\mathcal{N}(\mathbf{0},\Sigma_r)\\ &= \mathcal{N}(\mathbf{0},t_1\Sigma_1+\cdots+(t_r-t_{r-1})\Sigma_r),\end{aligned}$$

where the variance-covariance matrices are given by

$$\Sigma_j=\operatorname{diag}(\mathbf{p}_j)-\mathbf{p}_j^{\top}\mathbf{p}_j.$$

By a straightforward (linear) transformation consisting of taking partial sums, the random vector of increments $\Delta^{(n)}_{\lfloor \mathbf{t}n\rfloor}$ can be transformed into $\mathbf{K}^{(n)}_{\lfloor \mathbf{t}n\rfloor}$. This proves that $\mathbf{K}^{(n)}_{\lfloor \mathbf{t}n\rfloor}$ converges, after centering and rescaling, to a multivariate normal distribution.

The entries of the corresponding variance-covariance matrix can either be determined mechanically from the entries of $t_1\Sigma_1+\cdots+(t_r-t_{r-1})\Sigma_r$ by taking the partial summation into account, or alternatively, our observations concerning the martingale from Section A.1 can be used. In particular, using (11), we find, for fixed $s,t\in[0,1]$ with $s<t$, that

$$\begin{aligned}\operatorname{Cov}\Big(\frac{K^{(n)}_{\lfloor sn\rfloor}-\mathbb{E}K^{(n)}_{\lfloor sn\rfloor}}{\sqrt{n}},\frac{K^{(n)}_{\lfloor tn\rfloor}-\mathbb{E}K^{(n)}_{\lfloor tn\rfloor}}{\sqrt{n}}\Big) &= \frac{(n-\lfloor tn\rfloor)(n-\lfloor sn\rfloor)}{n}\mathbb{E}(Y^{(n)}_{\lfloor sn\rfloor}Y^{(n)}_{\lfloor tn\rfloor})\\ &= (n(1-t)(1-s)+O(1))\mathbb{E}({Y^{(n)}_{\lfloor sn\rfloor}}^2)\\ &= s^2(1-t)+O(n^{-1}),\end{aligned}$$

where we made use of the martingale property, and that the second moment of $Y^{(n)}_j$ is equal to the variance $n^{-2}j(j-1)/(n-j)$. Ultimately, this verifies (14) and thus completes the proof. ◀