La Matematica (2023) 2:861–892
https://doi.org/10.1007/s44007-023-00060-3

ORIGINAL RESEARCH ARTICLE

The Uncover Process for Random Labeled Trees

Benjamin Hackl[1,2,3] · Alois Panholzer[4] · Stephan Wagner[3]

Received: 30 December 2022 / Revised: 9 August 2023 / Accepted: 14 August 2023 /
Published online: 2 October 2023

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 work: 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; and lastly, the largest connected component, for which we also observe a phase transition.

Keywords Labeled tree · Uncover process · Functional central limit theorem · Limiting distribution · Phase transition

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 from 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

✉ Benjamin Hackl
math@benjamin-hackl.at

Alois Panholzer
alois.panholzer@tuwien.ac.at

Stephan Wagner
stephan.wagner@math.uu.se

[1] University of Graz, Graz, Austria

[2] University of Klagenfurt, Klagenfurt, Austria

[3] Uppsala University, Uppsala, Sweden

[4] TU Wien, Vienna, Austria

one by one in order of their labels. This yields a growing sequence of 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. See Fig. 1 for an illustration of this process on a tree with 100 vertices.

This model is motivated by stochastic models known as coalescent models for particle coalescence, most notably the additive and the multiplicative coalescent [3] and the Kingman coalescent [13]. To make the distinction between these classical coalescent models and our model more explicit, let us briefly revisit the additive coalescent model (see [2]) as a prototypical example. This model describes a Markov process on a state space consisting of tuples $(x_1, x_2, \ldots)$ 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" [6] and in a certain scheme for merging forests by uncovering one edge in every time step [20]. There is also a rich literature on random (edge) cutting of trees, starting with the work of Meir and Moon [17], see also [1, 7, 11, 15]. Fragmentation processes on trees have also been studied extensively in a continuous setting, see, e.g., [4, 18, 19, 25]. Lastly, the

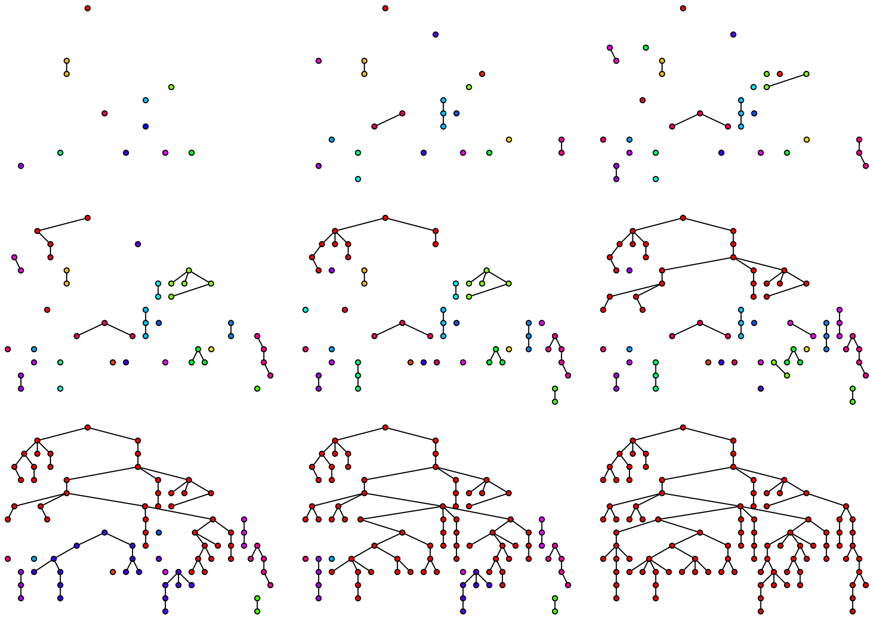

Fig. 1 A few snapshots of the *uncover process* applied to a random labeled tree of size 100. From left to right and top to bottom, there are 12, 23, 34, …, 89, and 100 uncovered vertices in the figures, respectively. Vertex labels are omitted for the sake of readability, and vertices are colored per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

special case of our model in which the underlying tree is always a path (with random labels) rather than a random labeled tree was considered by Janson in [12].

While the 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 [21], 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.

Overview. Different aspects of the uncover process on labeled trees are investigated in this work. In Sect. 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 3.

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 Sect. 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 7. 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 9, a characterization of the different limiting distributions for the root cluster size depending on the number of uncovered vertices.

Finally, in Sect. 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 functions 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

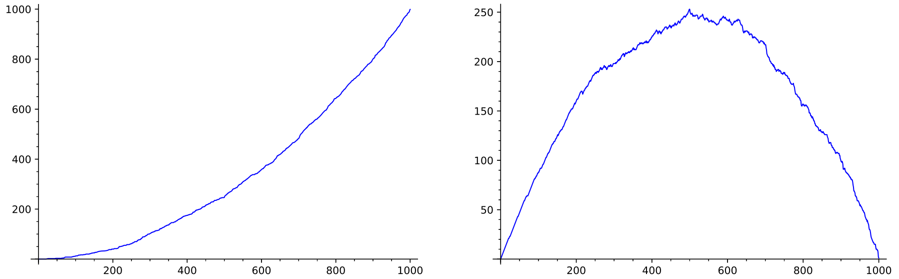

Fig. 2 Progression of the number of edges (left) and the number of connected components (right) when sequentially uncovering a random labeled tree on 1000 vertices

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 number of vertices in 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.

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 that is fully determined by its uncover sequence.

The following theorem provides an explicit formula for the (multivariate) generating function that tracks the number of edge increments over specified discrete time intervals of the uncover process.

Theorem 1 *Let r be a fixed positive integer with $1 \leq r < n - 1$, and let j_1, j_2, ..., j_r be positive integers with $1 < j_1 < j_2 < \cdots < j_r < n$. Additionally, let $j_0 = 1$. Then, the multivariate generating function tracking the number of uncovered edges when uncovering vertices in $[j_i + 1, j_{i+1}]$ for $0 \leq i < r$ in the edge uncover process is given by*

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

In other words, given a non-decreasing sequence of non-negative integers $0 = a_0 \leq a_1 \leq \cdots \leq a_r = n - 1$ with $a_i < j_i$, 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 $1 \leq i \leq r$.

Remark 2 By specifying the integers j_1, j_2, ..., j_r, the uncover process is effectively partitioned into intervals. This is also reflected by the quantities occurring in the product in (1): 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 Theorem 1 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); $E_n(z_1, \dots, z_r) = \sum_{|T|=n} w(T)$.

Following Martin and Reiner [16, Theorem 4] or Remmel and Williamson [22, Equation (8)], the generating function of the tree weights $w(T)$ has the explicit formula

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

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

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

In this form, the different values assumed by the sum when j moves through the ranges $1 < j \le j_1$, $j_1 < j \le j_2$, etc. can be determined directly. For some j with $j_{i-1} < j \le j_i$, the contribution to the product in (2) is

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

and for $j_r < j \le n - 1$ 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 (2) can be rewritten as the right-hand side of (1) and thus proves the lemma. □

With a 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.

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

Corollary 2 *In the setting of Theorem 1, 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*

$$(n-j_r)^{j_r-a_r-1} n^{n-j_r-1} \times \prod_{i=1}^{r} \Bigg(\sum_{h=0}^{a_i-a_{i-1}} \binom{j_{i-1}-a_{i-1}-1}{h} \binom{j_i - j_{i-1}}{a_i - a_{i-1} - h} (j_i - j_{i-1})^h j_i^{a_i-a_{i-1}-h} \Bigg). \tag{3}$$

Furthermore, there are

$$\prod_{i=1}^{n-2} \left(\binom{i-a_i-1}{a_{i+1}-a_i-1}(i+1) + \binom{i-a_i-1}{a_{i+1}-a_i} \right) \tag{4}$$

trees with a fully specified uncover sequence $(0, a_2, a_3, \ldots, a_{n-1}, n-1)$.

Proof The formulas follow from extracting the coefficient of $z_1^{a_1} z_2^{a_2-a_1} \cdots z_r^{a_r-a_{r-1}}$ from the corresponding generating function (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 $(K_j^{(n)})_{1\le j\le n}$ 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

$$\mathbb{E}K_j^{(n)} = (n-1)\frac{\binom{j}{2}}{\binom{n}{2}} = \frac{j(j-1)}{n}. \tag{5}$$

To motivate our investigations further, consider the illustrations in Fig. 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,

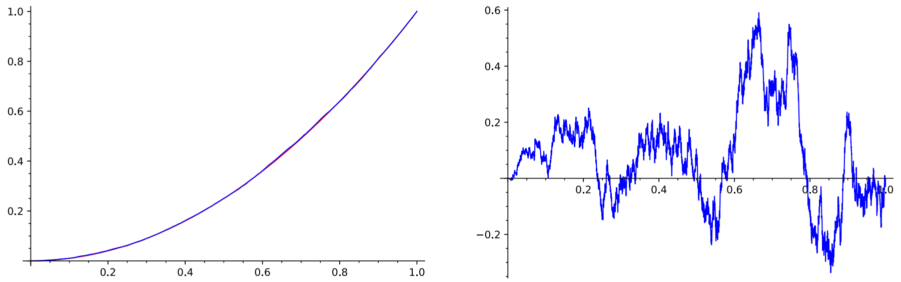

Fig. 3 A path of the rescaled stochastic process $(K^{(n)}_{\lfloor tn\rfloor}/(n-1))_{t\in[0,1]}$ (left-hand side) and the corresponding (rescaled) deviation $\big(\frac{\tilde K^{(n)}_t - t^2 n}{\sqrt{n}}\big)_{t\in[0,1]}$ for a random labeled tree of size $n = 10000$

see [14, Definition 21.8]. A Brownian bridge can then be defined by setting

$$B(t) = (1-t)W(t/(1-t)), \tag{6}$$

see, e.g., [23, 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 $K^{(n)}_0 = 0$ and introduce the linearly interpolated process $(\tilde K^{(n)}_t)_{t\in[0,1]}$, where

$$\tilde K^{(n)}_t := (1 + \lfloor tn\rfloor - tn)K^{(n)}_{\lfloor tn\rfloor} + (tn - \lfloor tn\rfloor)K^{(n)}_{\lceil tn\rceil}, \tag{7}$$

which by construction has continuous paths.

Theorem 3 *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*

$$Z^{\infty}(t) = (1-t)W\big(t^2/(1-t)\big). \tag{8}$$

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

$$\mathbb{E}Z^{\infty}(t) = 0, \quad \mathbb{V}Z^{\infty}(t) = t^2(1-t), \quad \text{and} \quad \operatorname{Cov}(Z^{\infty}(s), Z^{\infty}(t)) = s^2(1-t). \tag{9}$$

Remark 3 While the limiting process $(Z^{\infty}(t))_{t\in[0,1]}$ is not a Brownian bridge (the corresponding variances and covariances as given in (9) do not match), it is closely related. Comparing the characterization of $Z^{\infty}(t)$ in (8) to (6), 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 (cf. [14, Theorem 21.38]): the fact that the sequence of stochastic processes is *tight* on the one hand, and information on the finite-dimensional joint distributions of $(\tilde{K}_t^{(n)})_{t\in[0,1]}$ on the other hand.

By Prohorov's theorem ([14, Theorem 13.29]), tightness is equivalent to the sequence being weakly relatively sequentially compact. We prove that this is the case by checking Kolmogorov's criterion [14, Theorem 21.42] for which we have to verify that the family of initial distributions $(\tilde{Z}^{(n)}(0))_{n\in\mathbb{Z}_{\geq 0}}$ is tight and that the paths of $(Z^{(n)}(t))_{t\in[0,1]}$ cannot change too fast.

Let us begin with some probabilistic observations that will be very useful in the proof. Consider the generating function derived in (1). 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 $\mathbf{j} = (2, 3, \dots, n-1)$, is

$$P_n(z_2, \dots, z_{n-1}) = \prod_{i=2}^{n-1} \left(\frac{1}{n} + \frac{i}{n} z_i + \sum_{h=i+1}^{n-1} \frac{1}{n} z_h \right). \tag{10}$$

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 $j+1$ 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]

$$K_{j+1}^{(n)} = K_j^{(n)} + \operatorname{Ber}\Big(\frac{j+1}{n}\Big) + \operatorname{Bin}\Big(j-1-K_j^{(n)}, \frac{1}{n-j}\Big). \tag{11}$$

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.

We can actually use a similar approach to determine an explicit characterization of the distribution of $K_j^{(n)}$. Instead of constructing the probability generating function for a complete uncover sequence, we consider the probability generating function for $K_j^{(n)}$, obtained by normalizing the generating function from (1) for $r = 1$ and $j_1 = j$.

[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.

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

We find

$$P_n(z) = \frac{n^{n-j-1}(n-j+jz)^{j-1}}{n^{n-2}} = \Big(\frac{n-j}{n} + \frac{j}{n}z\Big)^{j-1},$$

which is exactly the probability generating function of a binomially distributed random variable. This proves $K_j^{(n)} \sim \mathrm{Bin}(j-1, j/n)$ and can, for example, be used to determine the variance of $K_j^{(n)}$ as

$$\mathbb{V}K_j^{(n)} = \frac{(n-j)(j-1)j}{n^2}. \tag{12}$$

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

$$Y_j^{(n)} := \frac{K_j^{(n)} - \frac{j(j-1)}{n}}{n-j}. \tag{13}$$

With the help of the recursive description in (11), we can show that $(Y_j^{(n)})_{1\le j\le n-1}$ is a martingale by computing

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

As an immediate consequence of the construction of $Y_j^{(n)}$ and (12) we find

$$\mathbb{V}Y_j^{(n)} = \frac{\mathbb{V}K_j^{(n)}}{(n-j)^2} = \frac{(j-1)j}{(n-j)n^2}.$$

Let us now consider the behavior of the finite-dimensional distributions.

Lemma 4 *Let r be a fixed positive integer, and let $\mathbf{t} = (t_1, \dots, t_r) \in (0,1)^r$ with $t_1 < t_2 < \cdots < t_r$. Then for $n \to \infty$, the random vector*

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

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

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

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

$$\mathbb{E}\mathbf{K}^{(n)}_{\lfloor \mathbf{t}n \rfloor} = n(t_1^2, t_2^2, \dots, t_r^2) + \mathcal{O}(1), \tag{14}$$

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

$$\sigma_{i,j} = \begin{cases} t_i^2(1-t_j) & \textit{if } i \le j, \\ t_j^2(1-t_i) & \textit{if } i > j. \end{cases} \tag{15}$$

Proof As a consequence of Theorem 1 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

$$\begin{aligned} P_n(z_1, z_2, \dots, z_r) &= \frac{E_n(z_1, z_2, \dots, z_r)}{n^{n-2}} \\ &= \prod_{i=1}^{r}\Big(1 - \frac{j_r}{n} + \frac{j_i}{n}z_i + \sum_{h=i+1}^{r} \frac{j_h - j_{h-1}}{n} z_h\Big)^{j_i - j_{i-1}}, \end{aligned} \tag{16}$$

where $j_0 = 1$ for the sake of convenience. Given that this probability generating function factors nicely, we could use a general result like the multidimensional quasi-power theorem (cf. [10]) to prove that the corresponding random vector

$$\Delta^{(n)}_{\mathbf{j}} = (K^{(n)}_{j_1}, K^{(n)}_{j_2} - K^{(n)}_{j_1}, \dots, K^{(n)}_{j_r} - K^{(n)}_{j_{r-1}})$$

converges, after suitable rescaling, to a multivariate Gaussian limiting distribution. However, there is a simple probabilistic argument: 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 $M_j \sim \text{Multi}(j_i - j_{i-1}, \mathbf{p}_i)$ where

$$\mathbf{p}_i = (p_{i,0}, p_{i,1}, \dots, p_{i,r}) \in [0,1]^r \quad \text{such that} \quad p_{i,h} = \begin{cases} 1 - t_r & \text{if } h = 0, \\ 0 & \text{if } 0 < h < i, \\ t_i & \text{if } h = i, \\ t_h - t_{h-1} & \text{otherwise.} \end{cases} \tag{17}$$

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

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

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 (16). 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 $1 \le i \le r$ denote the corresponding marginalized multinomial distributions instead.

For the sake of convenience, we make a slight adjustment: instead of fixing the integer vector $\mathbf{j} = (j_1, \dots, j_r)$, we fix $\mathbf{t} = (t_1, \dots, t_r)$ 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 $M \sim \operatorname{Multi}(n, \mathbf{p})$ converges, for $n \to \infty$ and after appropriate scaling, in distribution to a multivariate normal distribution,

$$\frac{M - n\mathbf{p}}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \operatorname{diag}(\mathbf{p}) - \mathbf{p}^\top \mathbf{p}). \tag{18}$$

As a consequence, we find that

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

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 $\Delta_{\lfloor \mathbf{t} n \rfloor}^{(n)}$ can be transformed into $\mathbf{K}_{\lfloor \mathbf{t} n \rfloor}^{(n)}$. This proves that $\mathbf{K}_{\lfloor \mathbf{t} n \rfloor}^{(n)}$ 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

$Y_j^{(n)}$ can be used. In particular, using (13), we find, for fixed $s, t \in [0, 1]$ with $s < t$, that

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

where we made use of the martingale property, and the fact that the second moment of $Y_j^{(n)}$ is equal to the variance $n^{-2}j(j-1)/(n-j)$. Ultimately, this verifies (15) and thus completes the proof. □

The last remaining piece required to prove that the sequence of processes $(Z^{(n)})_{n\geq 1}$ is tight is a bound on the growth of the corresponding paths.

Lemma 5 *There is a constant λ such that the following inequality holds for all $s, t \in [0, 1]$ and all $n \in \mathbb{Z}_{>0}$:*

$$\mathbb{E}\big(\big|Z^{(n)}(t) - Z^{(n)}(s)\big|^4\big) \leq \lambda\, |t-s|^2. \tag{19}$$

Proof Let us first consider the case where both $sn = \ell$ and $tn = m$ are integers. Assume that $\ell > m$. We can follow the argument in the proof of Lemma 4 to see that the random variable $K_\ell^{(n)} - K_m^{(n)}$ is distributed like the last component in the sum of two independent multinomial distributions, which gives us

$$K_\ell^{(n)} - K_m^{(n)} \sim \operatorname{Bin}\left(m-1, \frac{\ell-m}{n}\right) + \operatorname{Bin}\left(\ell-m, \frac{\ell}{n}\right).$$

Let us write $\operatorname{Bin}^*(n, p)$ for a centered binomial distribution, i.e., a binomial distribution $\operatorname{Bin}(n, p)$ with the mean np subtracted. Then we have

$$K_\ell^{(n)} - K_m^{(n)} - \frac{\ell^2 - m^2}{n} \sim \operatorname{Bin}^*\left(m-1, \frac{\ell-m}{n}\right) + \operatorname{Bin}^*\left(\ell-m, \frac{\ell}{n}\right) - \frac{\ell-m}{n}. \tag{20}$$

The fourth moment of a $\operatorname{Bin}^*(n, p)$-distributed random variable, which is the fourth centered moment of a $\operatorname{Bin}(n, p)$-distributed random variable, is

$$np(1-p)\big(1+(3n-6)p(1-p)\big) \leq np(1+3np).$$

Note that for the Bin^*-variables in (20), we have $\frac{(m-1)(\ell-m)}{n} \leq \ell - m$ and $\frac{(\ell-m)\ell}{n} \leq \ell - m$. Thus the fourth moments of the two centered binomial random variables are bounded above by

$$(\ell-m)(1+3(\ell-m)) \leq 4(\ell-m)^2.$$

So $K_\ell^{(n)} - K_m^{(n)} - \frac{\ell^2-m^2}{n}$ is the sum of three random variables (the third one actually constant) whose fourth moments are bounded above by $4(\ell-m)^2$, $4(\ell-m)^2$, and $(\ell-m)^4 n^{-4} \le (\ell-m)^2$, respectively. Applying the inequality $\mathbb{E}((X+Y+Z)^4) \le 27(\mathbb{E}(X^4)+\mathbb{E}(Y^4)+\mathbb{E}(Z^4))$, which is a simple consequence of Jensen's inequality, we get

$$\mathbb{E}\Big(K_\ell^{(n)} - K_m^{(n)} - \frac{\ell^2-m^2}{n}\Big)^4 \le 27\cdot(4+4+1)(\ell-m)^2 = 243(\ell-m)^2.$$

So if $tn=\ell$ and $sn=m$ are integers, we have

$$\begin{aligned}\mathbb{E}\big(\big|Z^{(n)}(t)-Z^{(n)}(s)\big|^4\big) &= \mathbb{E}\Bigg(\frac{K_\ell^{(n)}-\frac{\ell^2}{n}}{\sqrt{n}} - \frac{K_m^{(n)}-\frac{m^2}{n}}{\sqrt{n}}\Bigg)^4\\ &\le \frac{243(\ell-m)^2}{n^2} = 243(t-s)^2.\end{aligned}$$

Second, consider the case that tn and sn lie between two consecutive integers m and $m+1$: $\tilde{s}n = m \le sn \le tn \le m+1 = \tilde{t}n$. In this case, we can express the difference $Z^{(n)}(t)-Z^{(n)}(s)$ (using the linear interpolation in the definition of $\tilde{K}_t^{(n)}$) as

$$Z^{(n)}(t)-Z^{(n)}(s) = (t-s)n\big(Z^{(n)}(\tilde{t})-Z^{(n)}(\tilde{s})\big) + (t-s)\sqrt{n}(\tilde{t}-t-s+\tilde{s}).$$

The fourth moment of the first term is

$$\mathbb{E}\Big((t-s)n\big(Z^{(n)}(\tilde{t})-Z^{(n)}(\tilde{s})\big)\Big)^4 \le (t-s)^4 n^4\cdot 243(\tilde{t}-\tilde{s})^2 \le 243(t-s)^2$$

since $|t-s| \le |\tilde{t}-\tilde{s}| = \frac{1}{n}$. Likewise, the fourth power of the second term is easily seen to be bounded above by $(t-s)^2$. So the elementary inequality $\mathbb{E}((X+Y)^4) \le 8(\mathbb{E}(X^4)+\mathbb{E}(Y^4))$ yields

$$E\big(\big|Z^{(n)}(t)-Z^{(n)}(s)\big|^4\big) \le 8(243+1)(t-s)^2 = 1952(t-s)^2.$$

Finally, in the general case that t and s are arbitrary real numbers in the interval $[0,1]$ such that tn and sn do not lie between consecutive integers, we can write

$$\begin{aligned}Z^{(n)}(t)-Z^{(n)}(s) = \big(Z^{(n)}(t)-Z^{(n)}(u_1)\big) + \big(Z^{(n)}(u_1)-Z^{(n)}(u_2)\big)\\ +\big(Z^{(n)}(u_2)-Z^{(n)}(s)\big)\end{aligned}$$

for some real numbers u_1, u_2 with $s \le u_2 \le u_1 \le t$ such that $u_1 n$ and $u_2 n$ are integers and $tn \le u_1 n + 1$ as well as $sn \ge u_2 n - 1$. Combining the bounds from above and using again the inequality $\mathbb{E}((X+Y+Z)^4) \le 27(\mathbb{E}(X^4)+\mathbb{E}(Y^4)+\mathbb{E}(Z^4))$, we obtain

$$\begin{aligned}\mathbb{E}\big(\big|Z^{(n)}(t) - Z^{(n)}(s)\big|^4\big) &\le 27\big(1952(t-u_1)^2 + 243(u_1-u_2)^2 + 1952(u_2-s)^2\big)\\ &\le 27\cdot 1952(t-u_1+u_1-u_2+u_2-s)^2 = 52704(t-s)^2,\end{aligned}$$

completing the proof of the lemma with $\lambda = 52704$. □

All that remains now is to combine the two ingredients to prove our main result on the limiting process.

Proof of Theorem 3 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 [5, Theorems 7.1, 7.5]).

Tightness is implied (see [14, Theorems 13.29, 21.42]) by tightness of the initial distributions (which is satisfied in our case as every $\tilde{Z}^{(n)}(0)$ for $n \in \mathbb{Z}_{\ge 0}$ is degenerate and assumes value 0 with probability 1 as a consequence of the uncover process deterministically starting with no uncovered edges) and the moment bound in Lemma 5. The (limiting) behavior of the finite-dimensional distributions of the original process $(K^{(n)}_{\lfloor tn\rfloor})_{t\in[0,1]}$ is characterized by Lemma 4. This characterization carries over to the linearly interpolated process by an application of Slutsky's theorem [14, Theorem 13.18] after observing that

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

as a mechanical computation shows that $\mathbb{E}((K^{(n)}_{\lfloor tn\rfloor+1} - K^{(n)}_{\lfloor tn\rfloor})^2) = \mathcal{O}(1)$ (this also follows from Lemma 5).

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 4, we find for all $s, t \in [0,1]$ with $s < t$ that

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

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^{∞}. □

While we only needed to show tightness of the initial distributions of the processes $(Z^n(t))_{t\in[0,1]}$ to prove convergence to $Z^{\infty}(t)$, we can actually prove a much stronger result. A uniform bound (that implies tightness of $Z^{(n)}(t)$ for every fixed t) reads as follows.

Proposition 6 *For any real $C > 1$ and any positive integer n, the random variable $Z^{(n)}(t)$ satisfies the bound*

$$\mathbb{P}(\sup_{t\in[0,1]} |Z^{(n)}(t)| \geq C) \leq 4(C-1)^{-2}, \tag{21}$$

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.

Proof In order to obtain this 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 $j + \eta^2 \leq j + \eta = tn \leq n$. It follows that

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

so

$$\begin{aligned}
\mathbb{P}\big(\sup_{t\in[0,1]} |Z^{(n)}(t)| \geq C\big) &\leq \mathbb{P}\left(\sup_{0\leq j\leq n} \left| \frac{K_j^{(n)} - j(j-1)/n}{\sqrt{n}} \right| \geq C-1 \right) \\
&= \mathbb{P}\left(\sup_{1\leq j\leq n-1} \left| \frac{Y_j^{(n)}(n-j)}{\sqrt{n}} \right| \geq C-1 \right).
\end{aligned} \tag{22}$$

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 $(Y_j^{(n)})_{1\leq j\leq n-1}$ is a martingale, we can use Doob's maximal inequality [14, Theorem 11.2]. For any real $C > 0$ and any fixed integer k with $1 \leq k \leq n-1$, we have

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

With this, we have all required prerequisites to prove the bound. We partition the interval over which the supremum is taken in (22), 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

$$\begin{aligned}\mathbb{P}\left(\sup_{n-j\in I_i^{(n)}}\left|\frac{Y_j^{(n)}(n-j)}{\sqrt{n}}\right| \geq C-1\right) &\leq \mathbb{P}\left(\sup_{n-j\in I_i^{(n)}} |Y_j^{(n)}2^{-i+1}\sqrt{n}| \geq C-1\right)\\ &= \mathbb{P}\left(\sup_{n-j\in I_i^{(n)}} |Y_j^{(n)}| \geq \frac{2^{i-1}(C-1)}{\sqrt{n}}\right)\\ &\leq \frac{n}{2^{2i-2}(C-1)^2}\mathbb{V}(Y^{(n)}_{n-\lceil 2^{-i}n\rceil})\\ &\leq \frac{n}{2^{2i-2}(C-1)^2}\cdot\frac{2^i}{n} = \frac{4}{2^i(C-1)^2},\end{aligned}$$

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 (21) and therefore completes the proof. □

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 $R_n^{(k)}$ 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. [9, 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

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

Note that the first summand in the formal description of $\mathcal{F}$ corresponds to the case where the root vertex is covered, thus the size of the root cluster is zero.

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

$$\begin{aligned} F(z, u, v) &:= \sum_{T \in \mathcal{F}} \frac{z^{|T|} u^{\#U \text{ in } T} v^{\#V \text{ in } T}}{|T|!} \\ &= \sum_{n \geq 1} \sum_{0 \leq k \leq n} \sum_{m \geq 0} \frac{n^{n-1}}{n!} \binom{n}{k} \mathbb{P}\{R_n^{(k)} = m\} z^n u^k v^m, \\ G(z, u) &:= \sum_{T \in \mathcal{G}} \frac{z^{|T|} u^{\#U \text{ in } T}}{|T|!} = \sum_{n \geq 1} \sum_{0 \leq k \leq n} \frac{n^{n-1}}{n!} \binom{n}{k} z^n u^k, \end{aligned}$$

we obtain the characterizing equations

$$F = z e^G + zuv e^F, \qquad G = z(1+u) e^G. \tag{23}$$

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 (23), the following results on $R_n^{(k)}$ can be deduced.

Theorem 7 *The expectation $\mathbb{E}(R_n^{(k)})$ is, for $0 \leq k \leq n$ and $n \geq 1$, given by*

$$\mathbb{E}(R_n^{(k)}) = \sum_{j=1}^{k} \frac{j \, k^{\underline{j}}}{n^j}. \tag{24}$$

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

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

Proof After introducing

$$E := E(z,u) = \frac{\partial}{\partial v} F(z,u,v)\Big|_{v=1} = \sum_{n,k\geq 0} \frac{n^{n-1}}{n!}\binom{n}{k}\mathbb{E}(R_n^{(k)}) z^n u^k,$$

considering the partial derivative of (23) with respect to v easily yields

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

Extracting coefficients of E by an application of the Lagrange inversion formula (see, e.g., [9, Theorem A.2]) yields

$$\begin{aligned}[z^n]E &= \frac{1}{n}[G^{n-1}]\frac{u}{(1+u)(1-\frac{u}{1+u}G)^2} \cdot (1+u)^n e^{nG} \\ &= \sum_{j=0}^{n-1}(j+1)u^{j+1}(1+u)^{n-1-j} \cdot \frac{n^{n-2-j}}{(n-1-j)!},\end{aligned}$$

and further

$$[z^n u^k]E = \sum_{j=0}^{k-1}(j+1)\binom{n-1-j}{k-1-j}\frac{n^{n-2-j}}{(n-1-j)!}.$$

Using $\mathbb{E}(R_n^{(k)}) = \frac{[z^n u^k]E}{[z^n u^k]G} = \frac{n!\,[z^n u^k]E}{n^{n-1}\binom{n}{k}}$, we obtain the result stated in (24):

$$\begin{aligned}\mathbb{E}(R_n^{(k)}) &= n!\sum_{j=0}^{k-1}(j+1)\frac{\binom{n-1-j}{k-1-j}}{\binom{n}{k}} \cdot \frac{n^{-1-j}}{(n-1-j)!} = n!\sum_{j=0}^{k-1}\frac{j+1}{n^{1+j}(n-1-j)!} \cdot \frac{k^{\underline{j+1}}}{n^{\underline{j+1}}} \\ &= \sum_{j=1}^{k}\frac{j\,k^{\underline{j}}}{n^j}.\end{aligned}$$

In order to analyze the asymptotic behavior of $\mathbb{E}(R_n^{(k)})$, the following integral representation turns out to be advantageous,

$$\mathbb{E}(R_n^{(k)}) = \int_0^\infty (x-1)\,e^{-x}\left(1+\frac{x}{n}\right)^k \mathrm{d}x. \tag{25}$$

It might be considered as a variation of the integral representation of the so-called Ramanujan Q-function, $Q(n) = \sum_{j=0}^{n-1}\frac{(n-1)^{\underline{j}}}{n^j} = \int_0^\infty e^{-x}(1+\frac{x}{n})^{n-1}\mathrm{d}x$, which can be traced back to Ramanujan himself (see, e.g., [8]).

Equation (25) can be verified in a straightforward way by using the integral representation of the Gamma function:

$$\int_0^\infty (x-1)\, e^{-x}\Big(1+\frac{x}{n}\Big)^k \mathrm{d}x = \int_0^\infty (x-1)e^{-x}\sum_{j=0}^{k}\binom{k}{j}\frac{x^j}{n^j}\mathrm{d}x$$
$$= \sum_{j=0}^{k}\frac{\binom{k}{j}}{n^j}\left(\int_0^\infty e^{-x}x^{j+1}\mathrm{d}x - \int_0^\infty e^{-x}x^j\mathrm{d}x\right) = \sum_{j=0}^{k}\frac{\binom{k}{j}}{n^j}\big((j+1)!-j!\big)$$
$$= \sum_{j=0}^{k}\frac{k^{\underline{j}}\, j\, j!}{j!\, n^j} = \sum_{j=0}^{k}\frac{j\, k^{\underline{j}}}{n^j}.$$

In order to evaluate the integral (25) asymptotically, we first show that for the range $x \ge n^{\frac{1}{2}+\epsilon}$, with arbitrary but fixed $\epsilon > 0$, the contribution to the integral is exponentially small and thus negligible. Namely, when considering the integrand and setting $t = \frac{x}{n}$, we obtain the following estimate, uniformly for $k \in [0, n]$:

$$e^{-x}\Big(1+\frac{x}{n}\Big)^k \le e^{-x}\Big(1+\frac{x}{n}\Big)^n = e^{-x+n\ln(1+\frac{x}{n})} = e^{-n\big(t-\ln(1+t)\big)}.$$

Simple monotonicity considerations yield $t-\ln(1+t) \ge \frac{t}{4}$, for $t \ge 1$ (thus $x \ge n$), and $t-\ln(1+t) \ge \frac{t^2}{4}$, for $t \in [0, 1]$ (thus $x \in [0, n]$). Due to these estimates and by evaluating the resulting integral, we get the following bounds on the integral for the respective ranges:

$$\int_n^\infty (x-1)e^{-x}\Big(1+\frac{x}{n}\Big)^k \mathrm{d}x \le \int_n^\infty x e^{-\frac{x}{4}}\mathrm{d}x = 16\big(1+\frac{n}{4}\big)e^{-\frac{n}{4}},$$
$$\int_{n^{\frac{1}{2}+\epsilon}}^n (x-1)e^{-x}\Big(1+\frac{x}{n}\Big)^k \mathrm{d}x \le \int_{n^{\frac{1}{2}+\epsilon}}^\infty x e^{-\frac{x^2}{4n}}\mathrm{d}x = 2ne^{-\frac{1}{4}n^{2\epsilon}},$$

thus, by combining both cases, uniformly for $k \in [0, n]$ and $\epsilon \in (0, \frac{1}{2}]$:

$$\int_{n^{\frac{1}{2}+\epsilon}}^\infty (x-1)e^{-x}\Big(1+\frac{x}{n}\Big)^k \mathrm{d}x = \mathcal{O}\big(ne^{-\frac{1}{4}n^{2\epsilon}}\big).$$

Furthermore, we note that (roughly speaking) if k is sufficiently far away from n the integration range with negligible contribution can be extended. Namely, setting $\delta = 1-\frac{k}{n}$ and $x = nt$, we obtain for the integrand

$$e^{-x}\Big(1+\frac{x}{n}\Big)^k = e^{-n\big(t-\frac{k}{n}\ln(1+t)\big)} = e^{-n\big(t-(1-\delta)\ln(1+t)\big)} \le e^{-n\delta t} = e^{-\delta x} = e^{-(1-\frac{k}{n})x},$$

where we used the trivial estimate $\ln(1+t) \le t$, for $t \ge 0$. E.g., if $\delta \ge n^{-\frac{1}{4}}$, i.e., $k \le n - n^{\frac{3}{4}}$, one easily obtains from this estimate that the contribution to the integral from the range $x \geq n^{\frac{1}{4}+\epsilon}$ is asymptotically negligible:

$$\begin{aligned}\int_{n^{\frac{1}{4}+\epsilon}}^{n^{\frac{1}{2}+\epsilon}} (x-1)e^{-x}\Big(1+\frac{x}{n}\Big)^k \mathrm{d}x &\leq \int_{n^{\frac{1}{4}+\epsilon}}^{\infty} xe^{-(1-\frac{k}{n})x}\mathrm{d}x\\ &\leq \int_{n^{\frac{1}{4}+\epsilon}}^{\infty} xe^{-n^{-\frac{1}{4}}x}\mathrm{d}x = \mathcal{O}\big(ne^{-n^{\epsilon}}\big).\end{aligned}$$

To get the asymptotic expressions for the integral stated in the theorem, we will take into account the growth of k w.r.t. n, consider suitable expansions of the integrand for the range $x \leq n^{\frac{1}{2}+\epsilon}$ (or $x \leq n^{\frac{1}{4}+\epsilon}$, respectively) and evaluate the resulting integrals, where we apply the "tail exchange technique," i.e., we may extend the integration range to $x \geq 0$, since only asymptotically negligible contributions are added.

- k small or in the central region: assuming $k \leq n - n^{\frac{3}{4}}$, an expansion of the integrand for $x \leq n^{\frac{1}{4}+\epsilon}$ leads to the expansion (with a uniform bound in this range):

$$\begin{aligned}e^{-x}\Big(1+\frac{x}{n}\Big)^k &= e^{-x+k\ln(1+\frac{x}{n})} = e^{-x(1-\frac{k}{n})+\mathcal{O}(\frac{kx^2}{n^2})} = e^{-x(1-\frac{k}{n})}\cdot\Big(1+\mathcal{O}\big(\tfrac{kx^2}{n^2}\big)\Big)\\ &= e^{-x(1-\frac{k}{n})}\cdot\Big(1+\mathcal{O}\big(n^{-\frac{1}{2}+2\epsilon}\big)\Big),\end{aligned}$$

and thus to the following evaluation of the integral:

$$\begin{aligned}\int_0^{n^{\frac{1}{4}+\epsilon}} (x-1)e^{-x}\Big(1+\frac{x}{n}\Big)^k\mathrm{d}x &= \int_0^{n^{\frac{1}{4}+\epsilon}} (x-1)e^{-x(1-\frac{k}{n})}\mathrm{d}x\cdot\Big(1+\mathcal{O}\big(n^{-\frac{1}{2}+2\epsilon}\big)\Big)\\ &= \int_0^{\infty} (x-1)e^{-x(1-\frac{k}{n})}\mathrm{d}x\cdot\Big(1+\mathcal{O}\big(n^{-\frac{1}{2}+2\epsilon}\big)\Big)\\ &= \frac{\frac{k}{n}}{(1-\frac{k}{n})^2}\cdot\Big(1+\mathcal{O}\big(n^{-\frac{1}{2}+2\epsilon}\big)\Big),\end{aligned}$$

where we used the formula

$$\int_0^{\infty}(x-1)e^{-\nu x}\mathrm{d}x = \frac{1-\nu}{\nu^2}, \quad \text{for } \nu > 0. \tag{26}$$

Of course, this gives in particular

$$\mathbb{E}(R_n^{(k)}) \sim \begin{cases}\frac{k}{n}, & \text{for } k = o(n),\\ \frac{\alpha}{(1-\alpha)^2}, & \text{for } k \sim \alpha n, \text{ with } 0 < \alpha < 1.\end{cases}$$

- k subcritically large: in the following, we treat cases with $\frac{k}{n} \to 1$; there we have to distinguish according to the growth behavior of the difference $d = n - k$. First we examine the region $\sqrt{n} \ll d \ll n$, i.e., $d = o(n)$, but $d = \omega(\sqrt{n})$, for which we get the following expansion of the integrand (for $x \le n^{\frac{1}{2}+\epsilon}$):

$$\begin{aligned} e^{-x}\Big(1+\frac{x}{n}\Big)^k &= e^{-x+(n-d)\ln(1+\frac{x}{n})} = e^{-\frac{d}{n}x+\mathcal{O}(\frac{x^2}{n})+\mathcal{O}(\frac{x^3}{n^2})} \\ &= e^{-\frac{dx}{n}} \cdot \Big(1+\mathcal{O}\big(\tfrac{x^2}{n}\big)+\mathcal{O}\big(\tfrac{x^3}{n^2}\big)\Big). \end{aligned}$$

Considering the corresponding integral (and applying tail exchange), we obtain

$$\begin{aligned} \mathbb{E}(R_n^{(k)}) &= \int_0^\infty (x-1)e^{-\frac{dx}{n}}\,\mathrm{d}x \\ &\quad +\mathcal{O}\left(\frac{1}{n}\cdot\int_0^\infty (x-1)x^2e^{-\frac{dx}{n}}\,\mathrm{d}x\right)+\mathcal{O}\left(\frac{1}{n^2}\cdot\int_0^\infty (x-1)x^3e^{-\frac{dx}{n}}\,\mathrm{d}x\right). \end{aligned}$$

Using (26) and

$$\int_0^\infty (x-1)x^\ell e^{-\nu x}\,\mathrm{d}x = \mathcal{O}\big(\tfrac{1}{\nu^{\ell+2}}\big), \quad \text{for } \ell \ge 0 \text{ and } \nu \in (0,1),$$

we obtain the stated result:

$$\mathbb{E}(R_n^{(k)}) = \frac{n^2}{d^2}+\mathcal{O}\big(\tfrac{n}{d}\big)+\mathcal{O}\big(\tfrac{n^3}{d^4}\big) = \frac{n^2}{d^2}\cdot\Big(1+\mathcal{O}\big(\tfrac{d}{n}\big)+\mathcal{O}\big(\tfrac{n}{d^2}\big)\Big) \sim \frac{n^2}{d^2}.$$

- k critically large: for the case that the difference $d = n-k$ is of order $\Theta(\sqrt{n})$, we obtain the following expansion of the integrand:

$$e^{-x}\Big(1+\frac{x}{n}\Big)^k = e^{-\frac{dx}{n}-\frac{x^2}{2n}}\cdot\Big(1+\mathcal{O}\big(\tfrac{dx^2}{n^2}\big)+\mathcal{O}\big(\tfrac{x^3}{n^2}\big)\Big).$$

Thus, after completing the integrals occurring, we get

$$\begin{aligned} \mathbb{E}\big(R_n^{(k)}\big) &= \int_0^\infty (x-1)e^{-\frac{dx}{n}-\frac{x^2}{2n}}\,\mathrm{d}x \\ &\quad +\mathcal{O}\Big(\frac{d}{n^2}\cdot\int_0^\infty x^3e^{-\frac{dx}{n}-\frac{x^2}{2n}}\,\mathrm{d}x\Big)+\mathcal{O}\Big(\frac{1}{n^2}\cdot\int_0^\infty x^4e^{-\frac{dx}{n}-\frac{x^2}{2n}}\,\mathrm{d}x\Big). \end{aligned}$$

Since, for $\ell \ge 0$,

$$\int_0^\infty x^\ell e^{-\frac{dx}{n}-\frac{x^2}{2n}}\,\mathrm{d}x = \mathcal{O}\Big(\int_0^\infty x^\ell e^{-\frac{x^2}{2n}}\,\mathrm{d}x\Big) = \mathcal{O}\big(n^{\frac{\ell+1}{2}}\big),$$

this yields

$$\mathbb{E}\big(R_n^{(k)}\big) = \int_0^\infty x\,e^{-\frac{dx}{n}-\frac{x^2}{2n}}\,\mathrm{d}x + \mathcal{O}(\sqrt{n}).$$

Setting $d = c\sqrt{n}$ and applying the substitution $t = c + \frac{x}{\sqrt{n}}$, we evaluate the integral obtaining the stated result:

$$\int_0^\infty x\, e^{-\frac{dx}{n} - \frac{x^2}{2n}}\, \mathrm{d}x = n \int_c^\infty (t-c)\, e^{\frac{c^2}{2} - \frac{t^2}{2}}\, \mathrm{d}t = n\left(1 - ce^{\frac{c^2}{2}} \cdot \int_c^\infty e^{-\frac{t^2}{2}}\, \mathrm{d}t\right).$$

- k supercritically large: for $d = n - k = o(\sqrt{n})$, we get the expansion

$$e^{-x}\left(1 + \frac{x}{n}\right)^k = e^{-\frac{x^2}{2n}} \cdot \big(1 - \frac{dx}{n}\big) \cdot \Big(1 + \mathcal{O}\big(\tfrac{d^2x^2}{n^2}\big) + \mathcal{O}\big(\tfrac{x^3}{n^2}\big)\Big).$$

Computations analogous to the previous ones, using

$$\int_0^\infty x^\ell e^{-\frac{x^2}{2n}}\, \mathrm{d}x = 2^{\frac{\ell-1}{2}} \Gamma\big(\tfrac{\ell+1}{2}\big) \cdot n^{\frac{\ell+1}{2}}, \quad \text{for } \ell \geq 0,$$

lead to the stated result:

$$\begin{aligned}\mathbb{E}\big(R_n^{(k)}\big) &= \int_0^\infty xe^{-\frac{x^2}{2n}}\mathrm{d}x - \frac{d}{n}\int_0^\infty x^2 e^{-\frac{x^2}{2n}}\mathrm{d}x + \mathcal{O}(d^2) + \mathcal{O}(\sqrt{n}) \\ &= n - \frac{\sqrt{\pi}}{\sqrt{2}} d\sqrt{n} + \mathcal{O}(d^2) + \mathcal{O}(\sqrt{n}).\end{aligned}$$

□

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 (23), 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

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

Note that the second summand, $\frac{G}{1+u} = ze^G$, corresponds to the case where the root vertex is covered. Extracting coefficients via an application of the Lagrange inversion formula 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 $0 \leq m \leq k \leq n$ and $n \geq 1$):

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

From this formula, the probabilities $\mathbb{P}(R_n^{(k)} = m) = \frac{F_{n,k,m}}{n^{n-1}\binom{n}{k}}$ given in Theorem 9 can be obtained directly. We omit these straightforward, but somewhat lengthy computations, since in the following the results are deduced in a more general and elegant way.

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.

Lemma 8 *Let n and k be positive integers, and let $r_1, r_2, \dots, r_\ell$ be fixed positive integers with $r_1 + \cdots + r_\ell \le k$. Moreover, fix disjoint subsets $R_1, \dots, R_\ell$ of $[k]$ with $|R_i| = r_i$ for all i. The number of n-vertex labeled trees for which $R_1, \dots, R_\ell$ are components of the forest induced by the vertices with labels in $[k]$ is given by*

$$n^{n-k-1}\big(n - r_1 - \cdots - r_\ell\big)^{k-r_1-\cdots-r_\ell-1}(n-k)^\ell r_1^{r_1-1}\cdots r_\ell^{r_\ell-1}. \tag{27}$$

Proof We interpret each such tree T as a spanning tree of a complete graph K with n vertices. Set $r = r_1 + \cdots + r_\ell$. The vertices of K can be divided into the sets $R_1, \dots, R_\ell$, the remaining $k - r$ vertices in $\{1, 2, \dots, k\}$ forming a set Q, and the $n-k$ vertices with label greater than k forming a set S. Note first that the components induced by the sets $R_1, \dots, R_\ell$ can be chosen in $r_1^{r_1-2}\cdots r_\ell^{r_\ell-2}$ ways. If the vertex sets corresponding to $R_1, \dots, R_\ell$ are contracted to single vertices $v_1, \dots, v_\ell$, K becomes a multigraph K' with $n - r + \ell$ vertices, and the tree T becomes a spanning tree T' of K' upon contraction. Note that there are r_i edges from v_i to every other vertex in K', and that T' cannot contain any edges from v_i to another vertex in $\{v_1, \dots, v_\ell\} \cup Q$. Thus T' remains a spanning tree if all such edges are removed from K' to obtain a multigraph K''. Conversely, if we take an arbitrary spanning tree of K'' and replace the vertices $v_1, \dots, v_\ell$ by spanning trees of $R_1, \dots, R_\ell$, respectively, we obtain a labeled tree with n vertices that has the desired properties. It remains to count spanning trees of K'', which has an adjacency matrix of the block form

$$A = \begin{bmatrix} O & O & \mathbf{r}\mathbf{1}^T \\ O & E - I & E \\ \mathbf{1}\mathbf{r}^T & E & E - I \end{bmatrix}$$

Here, $\mathbf{1}$ denotes a (column) vector of 1s, $\mathbf{r}$ a (column) vector whose entries are $r_1, \dots, r_\ell$, O a matrix of 0s, E a matrix of 1s, and I an identity matrix. The blocks correspond to ℓ, $k - r$ and $n - k$ rows/columns, respectively. The number of spanning trees can now be determined by means of the matrix-tree theorem: the Laplacian matrix is given by

$$L = \begin{bmatrix} (n-k)D & O & -\mathbf{r}\mathbf{1}^T \\ O & (n-r)I - E & -E \\ -\mathbf{1}\mathbf{r}^T & -E & nI - E \end{bmatrix},$$

where D is a diagonal matrix with diagonal entries $r_1, \dots, r_\ell$. Our goal is to compute the determinant of L with the first row and column removed; let this matrix be L_1. If we subtract $\frac{1}{n-k}$ times the first $\ell - 1$ rows from all of the last $n - k$ rows of L_1, we obtain a matrix where all entries in the first $\ell - 1$ columns, except those in the diagonal, are 0. Thus the determinant is equal to the product of these diagonal entries

$r_2(n-k), \dots, r_\ell(n-k)$ times the determinant of a matrix of the block form

$$\begin{bmatrix} (n-r)I - E & -E \\ -E & nI - \left(1 + \frac{r-r_1}{n-k}\right)E \end{bmatrix},$$

where the blocks have length $k-r$ and $n-k$, respectively. This matrix has $n-r$ as an eigenvalue of multiplicity $k-r-1$, since subtracting $n-r$ times the identity yields a matrix with $k-r$ identical rows. For the same reason, n is an eigenvalue of multiplicity $n-k-1$. It remains to determine the remaining two eigenvalues. The corresponding eigenvectors can be constructed as follows: let the first $k-r$ entries (corresponding to the first block) be equal to a, and the remaining entries equal to b. It is easy to verify that this becomes an eigenvector for the eigenvalue λ if the simultaneous equations

$$\begin{aligned} (n-k)a - (n-k)b &= \lambda a, \\ -(k-r)a + (k-r+r_1)b &= \lambda b, \end{aligned}$$

are satisfied. The two solutions are the eigenvalues of the 2×2-coefficient matrix of this system, and their product is the determinant of this 2×2 matrix, which is

$$(n-k)(k-r+r_1) - (n-k)(k-r) = (n-k)r_1.$$

It finally follows that the determinant of L_1, thus the number of spanning trees of K'' is equal to

$$\begin{aligned} &r_2(n-k) \cdots r_\ell(n-k) \cdot (n-r)^{k-r-1} n^{n-k-1} (n-k) r_1 \\ &\quad = n^{n-k-1}(n-r)^{k-r-1}(n-k)^\ell r_1 \cdots r_\ell. \end{aligned}$$

Multiplying by $r_1^{r_1-2} \cdots r_\ell^{\ell-2}$ (the number of possibilities for the spanning trees induced in the components $R_1, \dots, R_\ell$), we obtain the desired formula. □

As a consequence of Lemma 8 for $\ell = 1$, the probability $\mathbb{P}(R_n^{(k)} = r)$ can be obtained by multiplying $n^{n-k-1}(n-r)^{k-r-1}(n-k)r^{r-1}$ with $r\binom{k}{r}$ (which gives the number of rooted labeled trees on n vertices whose root is contained in a cluster of size r among the first k uncovered vertices), and then normalizing by n^{n-1}, the number of labeled rooted trees on n vertices.

Theorem 9 *The exact distribution of $R_n^{(k)}$ is characterized by the following probability mass function (p.m.f.), which is given by the following formula for $0 \le m \le k \le n$ and $n \ge 1$ (and is equal to 0 otherwise):*

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

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

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

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

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

$$R_n^{(k)} \xrightarrow{d} R_\alpha, \quad \textit{where the discrete r.v. } R_\alpha \textit{ is characterized by its p.m.f.}$$
$$\mathbb{P}(R_\alpha = m) =: p_m = \begin{cases} 1-\alpha, & m = 0, \\ \frac{m^m}{m!}(1-\alpha)\alpha^m e^{-\alpha m}, & m \geq 1, \end{cases}$$

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

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

$$\left(\frac{d}{n}\right)^2 \cdot R_n^{(k)} \xrightarrow{d} \mathrm{G{\scriptsize AMMA}}\left(\frac{1}{2}, \frac{1}{2}\right),$$

where $\mathrm{G{\scriptsize AMMA}}(\frac{1}{2}, \frac{1}{2})$ 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 9 can be obtained in a rather straightforward way by applying Stirling's formula for the factorials after distinguishing several cases. □

Remark 4 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 9 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 $m \geq 1$.

Remark 5 Preliminary considerations indicate that the approaches used to characterize the distribution of $R_n^{(k)}$ could be extended, at least in principle, to obtain joint distributions of the size of the root cluster at several times during the uncover process. However, it seems that using them to deduce functional limit theorems for the size of the root cluster, e.g., in the critical region, might be a daunting task.

Remark 6 The distributions $R(c)$ and D occurring in the critical and supercritical region, resp., are related to the Lévy-distribution Lévy(γ), $\gamma > 0$, with density

$$f_\gamma(x) = \sqrt{\frac{\gamma}{2\pi}} \frac{e^{-\gamma/(2x)}}{x^{3/2}}, \quad x > 0.$$

Actually, D is a standard Levy-distributed random variable ($\gamma = 1$), whereas $R(c)$ can be obtained as the reciprocal of a shifted Lévy-distributed random variable $L \overset{d}{=} \text{Levy}(c^2)$:

$$R(c) \overset{d}{=} \frac{1}{1+L}.$$

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

$$\sum_{r=1}^{n} r \cdot X_{n,r}^{(k)}(T) = k. \tag{28}$$

Theorem 10 *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*

$$\mathbb{E}X_{n,r}^{(k)} = \binom{k}{r}\Big(\frac{r}{n}\Big)^{r-1}\Big(1-\frac{k}{n}\Big)\Big(1-\frac{r}{n}\Big)^{k-r-1}. \tag{29}$$

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 Lemma 8 and thus proves the theorem. □

In the spirit of the observation in (28), the formula in Lemma 8 provides a combinatorial proof for the following summation identity.

Corollary 11 *Let $n, k \in \mathbb{Z}_{\geq 0}$ with $0 \leq k \leq n$. Then, the identity*

$$\sum_{r=1}^{k}\binom{k}{r} r^r n^{n-k-1}(n-r)^{k-r-1}(n-k) = kn^{n-2} \tag{30}$$

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 7 Observe that the identity in (30) 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., [24].

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 12 *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 $d = o(\sqrt{n})$ (*supercritical case*), we have $c_{\max}^{(k)}(T_n)/n \xrightarrow{p} 1$.*

Informally, we can interpret this result as follows: as long as there are substantially more than an order of $\sqrt{n}$ many vertices left to be uncovered, then with high probability, there is no "giant" component (i.e., a component that contains at least a fixed percentage of all vertices). Otherwise, if there are substantially fewer than an order of $\sqrt{n}$ many vertices left to be uncovered, then with high probability there is one such "giant" component whose size is asymptotically equal to n.

Proof of Theorem 12 For the subcritical case, we use the expected root cluster size from Theorem 7. 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

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

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 supercritical case, we recall the corresponding case for the size of the root cluster from Theorem 7. 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. □

In the critical case where $n - k \sim c\sqrt{n}$ for a constant c, we are also able to characterize the behavior of $c_{\max}^{(k)}(T_n)/n$: this variable converges weakly to a continuous limiting distribution. In order to describe the distribution, we first require the following lemma.

Lemma 13 *Suppose that $n - k \sim c\sqrt{n}$, and fix $\alpha > 0$. The probability that the forest induced by the vertices with labels in $[k]$ of a uniformly random labeled tree T_n with n vertices contains two components, each with at least αn vertices, whose sizes are equal, goes to 0 as $n \to \infty$.*

Proof Suppose that there are two components with a vertices each, where $a \geq \alpha n$. By Lemma 8, the probability for this to happen is given by

$$P(a) = \frac{\binom{k}{a,a,k-2a} n^{n-k-1}(n-2a)^{k-2a-1}(n-k)^2 a^{2a-2}}{n^{n-2}},$$

provided that $k \geq 2a$ (otherwise, the probability is trivially 0). The initial multinomial coefficient gives the number of ways to choose the labels of the two components, the denominator is simply the total number of labeled trees. Our aim is to estimate this expression. The case $k = 2a$ is easy to deal with separately, so assume that $k > 2a$. Then by Stirling's formula, we have, for some constant C_1,

$$\begin{aligned} P(a) &\leq \frac{C_1 k^{k+1/2}}{a^{2a+1}(k-2a)^{k-2a+1/2}} \cdot \frac{n^{n-k-1}(n-2a)^{k-2a-1}(n-k)^2 a^{2a-2}}{n^{n-2}} \\ &= \frac{C_1 k^{1/2}(n-k)^2 n}{a^3(n-2a)(k-2a)^{1/2}} \Big(1 + \frac{n-k}{k}\Big)^{-k} \Big(1 + \frac{n-k}{k-2a}\Big)^{k-2a}. \end{aligned}$$

It is well known that $(1 + \beta/x)^x$ is increasing in x for fixed β. So if $k - 2a \leq n^{3/4}$, we have, using the assumption that $n - k \sim c\sqrt{n}$,

$$\begin{aligned} &\Big(1 + \frac{n-k}{k}\Big)^{-k} \Big(1 + \frac{n-k}{k-2a}\Big)^{k-2a} \\ &\leq \Big(1 + \frac{n-k}{k}\Big)^{-k} \Big(1 + \frac{n-k}{n^{3/4}}\Big)^{n^{3/4}} \\ &= \exp\Big(-k \log\Big(1 + \frac{n-k}{k}\Big) + n^{3/4} \log\Big(1 + \frac{n-k}{n^{3/4}}\Big)\Big) \\ &= \exp\Big(-k\Big(\frac{n-k}{k} + \mathcal{O}(n^{-1})\Big) + n^{3/4}\Big(\frac{n-k}{n^{3/4}} - \frac{(n-k)^2}{2n^{3/2}} + \mathcal{O}(n^{-3/4})\Big)\Big) \\ &= \exp\Big(-c^2 n^{1/4} + o(n^{1/4})\Big). \end{aligned}$$

In this case, $P(a)$ goes to 0 faster than any power of n. Otherwise, i.e., if $k - 2a > n^{3/4}$, we have $k - 2a \sim n - 2a$, and using the assumption that $a \geq \alpha n$ as well as the same Taylor expansion as above, we obtain

$$P(a) \leq \frac{C_2(n-k)^2}{n^{3/2}(n-2a)^{3/2}} \exp\Big(-\frac{(n-k)^2}{2(n-2a)}\Big)$$

for a constant C_2. We can rewrite this as

$$P(a) \leq \frac{C_2}{n^{3/2}(n-k)} f\Big(\frac{(n-k)^2}{n-2a}\Big)$$

with $f(x) = x^{3/2}e^{-x/2}$. Since this function is bounded, we have proven that $P(a) = \mathcal{O}(n^{-2})$, uniformly in a. Summing over all possible values of a, it follows that the probability in question is $\mathcal{O}(n^{-1})$. In particular, it goes to 0. □

Now we are able to prove the following description of the limiting distribution of $c_{\max}^{(k)}(T_n)/n$.

Theorem 14 *Suppose that $n-k \sim c\sqrt{n}$, and fix $\alpha > 0$. The probability that the forest induced by the vertices with labels in $[k]$ of a uniformly random labeled tree T_n with n vertices contains a component with at least αn vertices tends to*

$$\sum_{j\geq 1} \frac{(-1)^{j-1}c^j}{(2\pi)^{j/2}} \underset{\substack{\alpha\leq t_1<\cdots<t_j \\ \tau_j=t_1+\cdots+t_j<1}}{\int\cdots\int} \prod_{i=1}^{j} t_i^{-3/2}(1-\tau_j)^{-3/2} \exp\Big(-\frac{c^2\tau_j}{2(1-\tau_j)}\Big)\, \mathrm{d}t_1\cdots \mathrm{d}t_j$$

as $n\to\infty$.

Proof Let $r_1,\ldots,r_\ell$ be positive integers with $\alpha n \leq r_1 < \cdots < r_\ell$ and $r_1+\cdots+r_\ell \leq k$. By Lemma 8, the probability that the forest induced by vertices with labels in $[k]$ has components of sizes $r_1,\ldots,r_\ell$ is given by the following formula, with $r = r_1+\cdots+r_\ell$:

$$P(r_1,\ldots,r_\ell) = \frac{\binom{k}{r_1,\ldots,r_\ell,k-r} n^{n-k-1}(n-r)^{k-r-1}(n-k)^{\ell} r_1^{r_1-1}\cdots r_\ell^{r_\ell-1}}{n^{n-2}},$$

and the same argument as in Lemma 13 shows that this probability is $\mathcal{O}(n^{-\ell})$, uniformly in $r_1,\ldots,r_\ell$. Moreover, if we set $r_i = t_i n$, Stirling's formula gives us the following asymptotic formula for this probability after some manipulations: with $t = t_1+\cdots+t_\ell$, it is

$$P(r_1,\ldots,r_\ell) \sim \frac{c^\ell}{n^\ell(2\pi)^{\ell/2}} \prod_{i=1}^{\ell} t_i^{-3/2}(1-t)^{-3/2}\exp\Big(-\frac{c^2 t}{2(1-t)}\Big).$$

Moreover, by the inclusion–exclusion principle, the probability that there is at least one component of size at least αn is given by

$$\sum_{\alpha n\leq r_1\leq k} P(r_1) - \sum_{\substack{\alpha n\leq r_1<r_2 \\ r_1+r_2\leq k}} P(r_1,r_2) \\ +\cdots+(-1)^{j-1}\sum_{\substack{\alpha n\leq r_1<\cdots<r_j \\ r_1+\cdots+r_j\leq k}} P(r_1,\ldots,r_j)+\cdots+\mathcal{O}(n^{-1}).$$

The final error term takes the possibility into account that there are two components of the same size. The probability of this event is $\mathcal{O}(n^{-1})$ by Lemma 13. Note that we actually only need a finite number of terms, as the sums become empty for $j\alpha > 1$. If we plug in the asymptotic formula for $P(r_1,\ldots,r_\ell)$ and pass to the limit, the sums become integrals, and we obtain the desired formula. □

Acknowledgements We would like to thank Svante Janson for pointing out a gap in the proof of Theorem 3 in the extended abstract of this paper. Further we want to thank Markus Kuba for pointing out the relations to the Lévy-distribution for certain limiting distributions occurring in Theorem 9. Finally, we would like to thank the anonymous reviewers for their helpful comments.

Funding Open access funding provided by University of Graz. The author S. Wagner was supported by the Knut and Alice Wallenberg Foundation. An extended abstract of this paper appeared in the Proceedings of the 33rd International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms (AofA 2022).

Data availability Not applicable.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Addario-Berry, L., Broutin, N., Holmgren, C.: Cutting down trees with a Markov chainsaw. Ann. Appl. Probab. **24**(6), 2297–2339 (2014)
2. Aldous, D., Pitman, J.: The standard additive coalescent. Ann. Probab. **26**(4), 1703–1726 (1998)
3. Bertoin, J.: Random Fragmentation and Coagulation Processes. Cambridge Studies in Advanced Mathematics, vol. 102. Cambridge University Press, Cambridge (2006)
4. Berzunza Ojeda, G., Holmgren, C.: Invariance principle for fragmentation processes derived from conditioned stable Galton–Watson trees. arXiv Preprint (2010). arXiv:2010.07880
5. Billingsley, P.: Convergence of Probability Measures, 2nd edn. Wiley, New York (1999)
6. Chassaing, P., Louchard, G.: Phase transition for parking blocks, Brownian excursion and coalescence. Random Struct. Algorithms **21**(1), 76–119 (2002)
7. Fill, J.A., Kapur, N., Panholzer, A.: Destruction of very simple trees. Algorithmica **46**(3–4), 345–366 (2006)
8. Flajolet, P., Grabner, P.J., Kirschenhofer, P., Prodinger, H.: On Ramanujan's Q-function. J. Comput. Appl. Math. **58**(1), 103–116 (1995)
9. Flajolet, P., Sedgewick, R.: Analytic Combinatorics. Cambridge University Press, Cambridge (2009)
10. Heuberger, C., Kropf, S.: Higher dimensional quasi-power theorem and Berry–Esseen inequality. Monatsh. Math. **187**(2), 293–314 (2018)
11. Janson, S.: Random cutting and records in deterministic and random trees. Random Struct. Algorithms **29**(2), 139–179 (2006)
12. Janson, S.: Sorting using complete subintervals and the maximum number of runs in a randomly evolving sequence. Ann. Combin. **12**(4), 417–447 (2009)
13. Kingman, J.F.C.: The coalescent. Stoch. Process. Appl. **13**(3), 235–248 (1982)
14. Klenke, A.: Probability Theory—A Comprehensive Course, 3rd edn. Universitext. Springer, Cham (2020)
15. Kuba, M., Panholzer, A.: Isolating a leaf in rooted trees via random cuttings. Ann. Combin. **12**(1), 81–99 (2008)

16. Martin, J.L., Reiner, V.: Factorization of some weighted spanning tree enumerators. J. Combin. Theory Ser. A **104**(2), 287–300 (2003)
17. Meir, A., Moon, J.W.: Cutting down random trees. J. Austral. Math. Soc. **11**, 313–324 (1970)
18. Miermont, G.: Self-similar fragmentations derived from the stable tree. I. Splitting at heights. Probab. Theory Relat. Fields **127**(3), 423–454 (2003)
19. Miermont, G.: Self-similar fragmentations derived from the stable tree. II. Splitting at nodes. Probab. Theory Relat. Fields **131**(3), 341–375 (2005)
20. Pitman, J.: Coalescent random forests. J. Combin. Theory Ser. A **85**(2), 165–193 (1999)
21. Pitman, J.: Coalescents with multiple collisions. Ann. Probab. **27**(4), 1870–1902 (1999)
22. Remmel, J.B., Williamson, S.G.: Spanning trees and function classes. Electron. J. Combin. **9**(1), Research Paper 34, 24 (2002)
23. Revuz, D., Yor, M.: Continuous Martingales and Brownian Motion. Grundlehren der mathematischen Wissenschaften, vol. 293, 3rd edn. Springer, Berlin (1999)
24. Riordan, J.: Combinatorial Identities. Wiley, New York (1968)
25. Thévenin, P.: A geometric representation of fragmentation processes on stable trees. Ann. Probab. **49**(5), 2416–2476 (2021)