arXiv:2301.00664v1 [math.PR] 2 Jan 2023

随机标号树的揭示过程

BENJAMIN HACKL、ALOIS PANHOLZER 和 STEPHAN WAGNER

摘要。我们研究按照顶点标签揭示随机标号树顶点的过程。首先，均匀随机生成一棵含有 n 个顶点的标号树。随后，按照标签顺序逐一揭示顶点。每揭示一个新顶点，也同时揭示它与先前已揭示顶点之间的所有边。由此得到一列不断增长的森林。本文研究了这一过程的三个特定方面：首先是边的数量，我们证明经过适当缩放后，它会收敛到一种类似布朗桥的随机过程。其次是固定顶点所属的连通分量，我们识别出不同的阶段，并确定各阶段的极限分布。最后是最大的连通分量，我们也观察到了相变。

1. 引言

我们研究揭示随机树顶点的过程：首先从规模为 n （即含有 n 个顶点）的 n^{n-2} 棵无根标号树或 n^{n-1} 棵有根无序标号树中，均匀随机选取一棵，然后按照标签顺序逐一揭示顶点。由此得到一列由已揭示顶点诱导、不断增长的森林；我们关注这些森林从揭示第一个顶点到所有顶点均被揭示这一过程中的演变。

该模型的灵感来源于称为粒子聚并合并模型的随机模型，尤其是加性聚并和乘性聚并 [3]以及 Kingman凝聚 [13]。为了更明确地区分这些经典合并模型与我们的模型，我们简要回顾一下加性聚并模型（见 [2]），作为一个典型示例。该模型描述了一个定义在状态空间上的马尔可夫过程，其状态空间由元组 $(x_1, x_2, \ldots)$ 构成，其中 $x_1 \geq x_2 \geq \cdots \geq 0$ 和 $\sum_{i\geq 0} x_i = 1$，该模型描述单位质量碎裂成若干簇的过程

(Benjamin Hackl) 格拉茨大学，奥地利；克拉根福大学，奥地利；乌普萨拉大学，瑞典（Alois Panholzer）维也纳工业大学，奥地利（Stephan Wagner）乌普萨拉大学，瑞典电子邮箱地址：math@benjamin-hackl.at, alois.panholzer@tuwien.ac.at, stephan.wagner@math.uu.se。第三位作者获得了克努特与爱丽丝·瓦伦堡基金会的资助。本文的扩展摘要曾发表于第33届算法分析概率、组合与渐近方法国际会议论文集（AofA 2022）[9]。

，每个簇的质量为 x_i。质量分别为 x_i 和 x_j 的簇对随后以速率 $x_i + x_j$合并成一个质量为 $x_i + x_j$ 的新簇。在该过程对应的离散时间版本中，每个时间步恰好合并两个簇。这种离散加性凝聚模型出现在各种组合情境中，例如，与“线性探测散列” [6] 相关的停车方案中停车区块的演变，以及一种每个时间步通过揭开一条边来合并森林的方案 [20]。关于树的随机（边）切割，也有大量研究，始于 Meir 和 Moon 的工作 [17], ，另见 [1, 7, 11, 15]。树上的碎裂过程也已在连续情境下得到广泛研究，参见例如 [4, 18, 19, 25]。最后，Janson 在 [12]中研究了我们模型的一个特殊情形：底层树始终是一条路径（带有随机标号），而非随机标号树。

虽然逐步揭示边的加性聚并过程在理念上与我们的顶点揭露模型十分相近，但其底层过程却有着根本差异：这些经典聚并模型依赖于这样一个事实，即每个时间步恰好合并两个簇，而我们的模型并非如此。揭示一个新顶点时，揭示的边数可能多寡不定（也可能一条也没有）。还有一些聚并模型，例如 Λ-聚并过程，它是 Kingman凝聚 [21], 的一种推广，允许合并两个以上的簇——然而，据我们所知，目前还没有任何已知的聚并模型能够刻画顶点揭示过程的行为。

概述。本文研究了标号树上揭示过程的不同方面。在第 2 节中，我们研究由未揭示边数量构成的随机过程。相应的主要结果——对该过程及其极限行为的完整刻画——见定理 3。

第3节和第4节都关注簇大小，即整个过程中形成的连通分量的大小。具体而言，在第3节中，我们将注意力转向有根标号树，研究包含某个固定顶点的连通分量的行为。定理7分析了根簇的期望大小。此外，我们证明，根簇大小为给定值的有根树的数量可由一个相当简单的计数公式给出；这一公式进而体现于定理 9中，该定理根据未揭示顶点的数量，对根簇大小的不同极限分布作出刻画。

最后，在第4节中，我们利用关于根簇的结果，得出关于树中最大簇大小的结论。

记号。本文中，我们用 $[n] = \{1, \dots, n\}$ 和 $[k, \ell] = \{k, k+1, \dots, \ell\}$ 表示离散区间，用 $x^j = x(x-1) \cdots (x-j+1)$表示下降阶乘。分别用 $\lfloor x \rfloor$ 和 $\lceil x \rceil$表示向下取整函数和向上取整函数。此外，我们分别用 $\mathcal{T}$ 和 $\mathcal{T}^\bullet$ 表示标号树和有根标号树的组合类，并用 $\mathcal{T}_n$ 和 $\mathcal{T}_n^\bullet$ 表示大小为

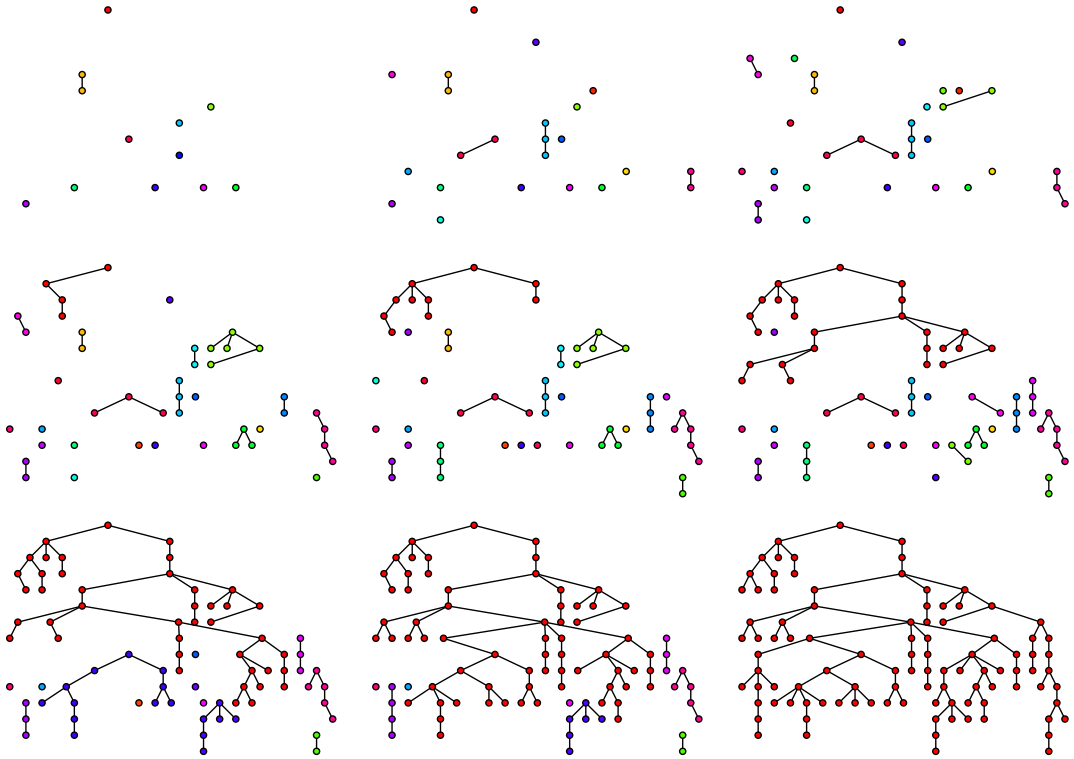

图 1. 将揭示过程 应用于一棵规模为 100 的随机标号树时的若干个快照。从左到右、从上到下，各图中分别有 12、23、34、.....、89 和 100 个未揭示顶点。为便于阅读，省略了顶点标签，并按连通分量为顶点着色。

n的标号树和有根标号树类，即含有 n 个顶点。最后，我们用 $X_n \xrightarrow{d} X$ 和$X_n \xrightarrow{p} X$ 分别表示随机变量序列（r.v.）$(X_n)_{n\geq 0}$ 依分布收敛于随机变量 X以及依概率收敛于该随机变量。

2. 未揭示边的数量

本节主要关注揭示过程中未揭示边数量的变化。我们首先为这一数量引入一个形式参数。

定义 1. 设 T 是一棵顶点集为 $V(T) = [n]$ 的标号树。对于 $1 \leq j \leq n$，我们用 $k_j(T) := \|T[1, 2, \ldots, j]\|$ 表示由 $[j]$ 中的顶点诱导出的 T 子图中的边数。我们将序列 $(k_j(T))_{1\leq j\leq n}$ 称为（边）揭露序列。

我们先给出一些简单的观察。首先，对于任意一棵阶为 n的标号树，都有 $k_1(T) = 0$，以及 $k_n(T) = n - 1$。其次，由于树的任意诱导子图都是森林，且森林具有一个基本性质：边数与连通分量数之和等于森林的阶，因此我们发现，$j - k_j(T)$ 表示揭露 T 的前 j 个顶点后连通分量的数量。图2展示了在一棵从1000个顶点的标号树中随机选取的树里，$1 \le j \le 1000$ 的边数和连通分量数的变化过程。

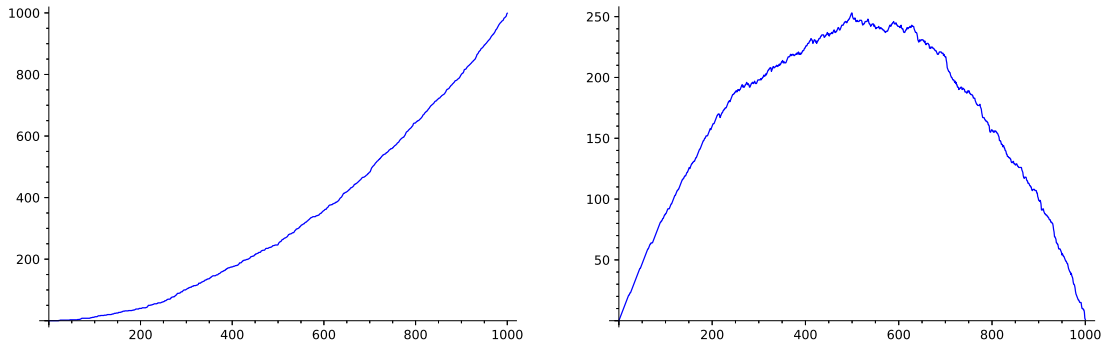

图2. 按顺序揭示一棵含有1000个顶点的随机标号树时，边数（左）和连通分量数（右）的变化。

此外，树的前 j 个顶点所诱导的子图是森林这一事实，也给出了对所有 $1 \le j \le n - 1$都成立的紧确界 $0 \le k_j(T) \le j - 1$。下界由中心顶点为 n的星形树取得，上界则由（线性有序的）路径取得。我们还可以观察到，一旦 $k_{n-1}(T) > 0$，最后一次揭示步骤中添加的边集就不再是唯一确定的。因此，中心顶点为 n 的星形树是唯一一棵可由其揭露序列完全确定的树。

下面的定理分别给出了具有部分指定和完全指定揭露序列的树的数量的显式计数公式。

定理1. 设 r 是满足 $1 \le r < n - 1$ 的固定正整数，设 j_1、j_2、……、j_r 是满足 $1 < j_1 < j_2 < \cdots < j_r < n$ 的正整数，设 a_1、a_2、……、a_r 是满足对所有 $1 \le i \le r$ 均有 $a_i \le j_i - 1$ 的非负整数非降序列。此外，设 $j_0 = 1$ 且 $a_0 = 0$。那么，满足对所有 $1 \le i \le r$ 均有 $k_{j_i}(T) = a_i$ 的 T 阶有根标号树的数量为

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

此外，还有

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

具有完全指定揭边序列 $(0, a_2, a_3, \ldots, a_{n-1}, n-1)$ 的树。

我们首先推导一个有用的辅助结果，即相应（多元）生成函数的显式公式。随后，通过提取适当的系数即可得到计数公式。

引理 2. 在定理*1*的设定下，边揭开过程中增量的多元生成函数为

$$E_n(z_1, z_2, \ldots, z_r) = n^{n-j_r-1}\prod_{i=1}^{r}\Big(n-j_r+j_i z_i+\sum_{h=i+1}^{r}(j_h-j_{h-1})z_h\Big)^{j_i-j_{i-1}}. \tag{3}$$

换言之，$E_n(z_1, \ldots, z_r)$ 的展开式中单项式 $z_1^{a_1}z_2^{a_2-a_1}\ldots z_r^{a_r-a_{r-1}}$的系数，等于阶为 n、满足对所有 $1 \le i \le r$ 均有 $k_{j_i}(T) = a_i$ 的 T棵标号树的数量。

注 2. 指定整数 $j_1, j_2, \ldots, j_r$ 后，揭示过程便被有效地划分为若干区间。这也反映在式 (3) 乘积中的各个量上：差值 $j_i - j_{i-1}$ 对应第 i 个区间中揭示的顶点数，j_i 表示截至第 i 个区间为止揭示的顶点总数，而 $n - j_r$ 对应最后一个区间中揭示的顶点数。

引理 *2* 的证明。我们首先注意到，当该过程揭示标签为 j 的顶点时，所有连接到标签小于 j 的相邻顶点的边也都会被揭示。为了确定边增量的生成函数，我们给连接顶点 i 和顶点 j 且具有 $i < j$ 的边赋予权重 $x_i y_j$，然后考虑树权重 $w(T)$ 的生成函数（其定义为边权重的乘积）；$E_n(z_1、\ldots\ldots、z_r) = \sum_{|T|=n} w(T)$。

根据 Martin 和 Reiner [16, 定理 4] 或 Remmel 和 Williamson [22,公式 (8)]，，树权重 $w(T)$ 的生成函数有如下显式公式

$$\sum_{|T|=n} w(T) = x_1 y_n \prod_{j=2}^{n-1}\Big(\sum_{i=1}^{n} x_{\min(i,j)} y_{\max(i,j)}\Big). \tag{4}$$

正如最初观察到的，由 $k_{j_i}(T)$ 计数的边，恰好是那些对某个 $\ell \le j_i$ 诱导出因子 y_ℓ 的边。因此，我们作如下替换：$x_\ell = 1$ 对所有 ℓ，$y_\ell = z_i$ 当且仅当 $j_{i-1} < \ell \le j_i$（其中[1] $j_0 = 1$），以及 $y_\ell = 1$ 如果

[1]注意到 y_1 不会出现，因为每条边的两个端点中至少有一个标签大于 1。

$\ell > j_r$。为处理关于 $y_{\max(i,j)}$ 的求和，注意到我们可以将其改写为

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

在这种形式下，当 j 在范围 $1 < j \le j_1$、$j_1 < j \le j_2$ 等中取值时，总和所取的不同值可以直接确定。对于某个满足 $j_{i-1} < j \le j_i$ 的 j，其对乘积 (4) 的贡献为

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

并且当 $j_r < j \le n-1$ 时，所有 y 变量都替换为 1，因此对乘积的贡献是因子 n。结合这两点可知，(4) 的右端可以改写为 (3) 的右端，从而证明该引理。□

有了揭示过程中边增量的生成函数的显式公式，计数具有给定（部分）揭露序列的树的数量便可直接得到。

定理*1*的证明。 接下来只需提取 $z_1^{a_1} z_2^{a_2-a_1} \cdots z_r^{a_r-a_{r-1}}$ 的系数，具体步骤是从 z_1 开始逐步进行。□

2.1. 深入考察随机过程。 边增量生成函数的公式格外优美，可用于更详细地研究描述未揭示边数量的随机过程。令随机变量序列 $(K_j^{(n)})_{1\le j\le n}$ 表示一个离散随机过程，用以模拟在大小为 n 的随机标号树中随机均匀选取的树，揭示前 j 个顶点后未揭示边的数量。利用一个简单的论证即可确定未揭示边的期望数量：有 j 个未揭示顶点时，$\binom{n}{2}$ 个可能的边位置中已有 $\binom{j}{2}$ 个被揭示。由于对称性以及标号树是均匀选取的，每个位置放置 $n-1$ 条边之一的可能性都相同，因此我们得到

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

为了进一步说明我们研究的动机，请看图 3 中的示意图。重新缩放后的偏离均值看起来很像一种称为布朗桥的随机过程。

为正式定义这一过程，首先回顾一下，维纳过程（$W(t)$)$_{t\in[0,1]}$ 是满足以下条件的唯一随机过程：

- $W(0) = 0$,
- W 具有独立、平稳的增量，

对所有 $t > 0$，均有 • $W(t) \sim \mathcal{N}(0, t)$，并且

- $t \mapsto W(t)$ 几乎必然连续，

参见 [14, 定义 21。8]随后可以通过设定来定义布朗桥

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

例如参见 [23, 练习 3。10]“桥”这一术语源于这一事实：我们有$B(0) = B(1) = 0$。

虽然均值的（标准化）偏差看起来似乎会收敛到布朗桥，但我们将证明，实际上只有几乎 如此。以下定理刻画了该随机过程。出于技术上的考虑，我们令 $K_0^{(n)}$ =0，并引入线性插值过程（$\tilde{K}_t^{(n)}$ ）$_{t\in[0,1]}$，其中

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

根据构造，该过程具有连续路径。

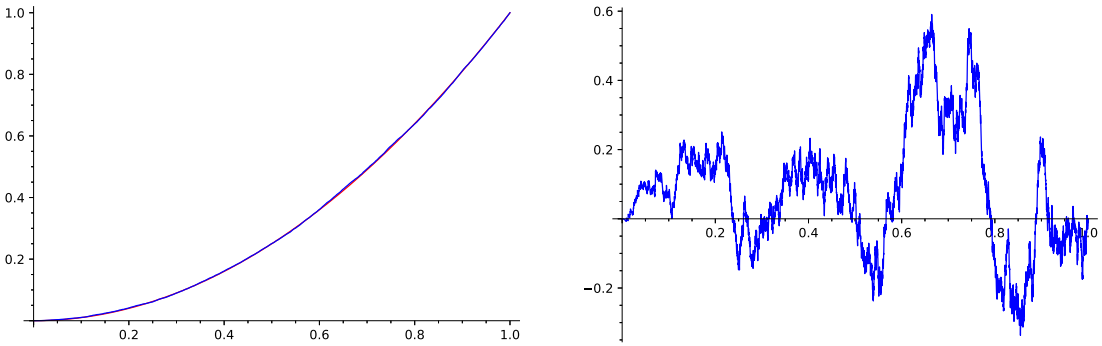

图 3. 对于大小为 $n = 10000$ 的随机标号树，重缩放随机过程（$K_{\lfloor tn \rfloor}^{(n)}/(n-1)$）$_{t\in[0,1]}$ 的一条路径（左侧）及其对应的（重缩放）偏差$\left(\frac{\tilde{K}_t^{(n)} - t^2 n}{\sqrt{n}}\right)_{t\in[0,1]}$。

定理 3. 令$(Z^{(n)}(t))_{t\in[0,1]}$ 为对线性插值过程$(\tilde{K}_t^{(n)})_{t\in[0,1]}$ 进行中心化和重缩放后得到的连续随机过程，其形式为

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

并令$(W(t))_{t\in[0,1]}$ 为标准维纳过程。那么，对于 $n \to \infty$，重缩放后的过程相对于上确界范数在 $C([0,1])$ 上弱收敛到极限过程 $Z^{\infty}(t)$ ，该极限过程由下式给出

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

此外，对于 $s, t \in [0,1]$ 且满足 $s < t$，极限过程满足

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

注释 3。虽然极限过程 $(Z^{\infty}(t))_{t\in[0,1]}$ 不是布朗桥（其方差和协方差与式 (9) 中给出的不一致），但二者密切相关。将式 (8) 中 $Z^{\infty}(t)$ 的刻画与式 (6) 比较可知，这两个过程唯一的区别在于维纳过程自变量的分子平方项。

要证明这一结果（参见 [14, 定理21.38]），需要两个主要要素：一方面，随机过程序列具有紧性；另一方面，需要了解$(\tilde{K}_t^{(n)})_{t\in[0,1]}$ 的有限维联合分布。

根据普罗霍罗夫定理（[14, 定理13.29]），紧性等价于该序列弱相对序列紧致。我们通过检验柯尔莫戈洛夫判据 [14, 定理21.42] 来证明这一点；为此，我们必须验证初始分布族 $(\tilde{Z}^{(n)}(0))_{n\in\mathbb{Z}_{\geq 0}}$ 是紧的，并且 $(Z^{(n)}(t))_{t\in[0,1]}$ 的路径变化不会太快。

初始分布的紧性相当容易得到（因为我们的过程在开始时是确定性的），但我们甚至可以推导出一个强得多的统一界。先重新考察式（3）。根据凯莱著名的树计数公式，当我们选择整数向量为$\mathbf{j}=(2,3,\ldots\ldots,n-1)$时，完整揭露序列对应的概率生成函数为

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

这表明，可以用 $n-2$ 个独立随机变量来对该过程建模，每个变量表示一个边增量[2]。因式分解表明，第 j 个增量（对应于含有 $i=j+1$ 的因子）在标签为 $j+1$ 的顶点尚未被揭示时，以 $(j+1)/n$ 的概率发生；而每当任意后续顶点尚未被揭示时，则以 $1/n$ 的概率发生。这个概率视角可用于对未揭示边的数量作递归刻画，即[3]

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

伯努利变量刻画的是在揭示标签为 $j+1$ 的顶点时，第 j 个边增量被添加的概率；二项变量则刻画所有其余尚未揭示的边增量。

[2]这里我们明确对边增量而非边进行建模，因为采用这种方法后，我们不必关心正在揭示的是哪一条边。我们的模型只明确刻画未揭示边数量的变化。[3]我们稍微滥用了记号：严格来说，我们需要引入辅助变量，使其分别服从指定的二项分布和伯努利分布。

现在，我们通过定义来考虑该过程 $(K_j^{(n)})_{1\le j\le n}$ 的中心化和重新缩放版本

$$Y_j^{(n)} := \frac{K_j^{(n)} - \frac{j(j-1)}{n}}{n-j}. \tag{12}$$

借助式 (11) 中的递归描述，我们可以通过计算证明 $(Y_j^{(n)})_{1\le j\le n-1}$是一个鞅

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

我们还可以给出 $Y_k^{(n)}$ 的方差的显式表达式：回顾式 (12)，有 $\mathbb{V}Y_k^{(n)} = (n-k)^{-2}\mathbb{V}K_k^{(n)}$。然后，利用式 (11) 以及全方差公式和全期望公式，我们得到递推关系

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

对于 $1 \le k < n-1$ 和 $\mathbb{V}K_1^{(n)} = 0$。由此可得

$$\mathbb{V}K_k^{(n)} = \sum_{j=1}^{k-1} \Big(\frac{n-k}{n-j-1}\Big)^2 \frac{(n-j-1)(2n-j-1)j}{(n-j)n^2} = \frac{k(k-1)(n-k)}{n^2}, \tag{13}$$

w这里的求和可以借助部分分式和裂项相消法求出 g.

现在可以将一致界（这也意味着，对于每个固定的 t， $Z^{(n)}(t)$ 具有紧性）表述如下。

引理 4。 对于任意实数 $C > 1$ 和任意正整数 n，随机变量$Z^{(n)}(t)$满足以下界

$$\mathbb{P}(\sup_{t\in[0,1]} |Z^{(n)}(t)| \ge C) \le 4(C-1)^{-2}, \tag{14}$$

因此，对于 $C \to \infty$，该过程的绝对值超过 C 的概率关于 n 一致收敛到 0。

证明。 为了得到这一条件，我们首先证明它可以归结为上一节中鞅的一个不等式。为此，我们写作

$tn = j + \eta$，其中 $j \in \mathbb{Z}$ 和 $\eta \in [0, 1)$。简单计算可得

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

最终的分数不超过 1，因为 $j + \eta^2 \le j + \eta = tn \le n$。由此可得

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

so

$$\begin{aligned} \mathbb{P}\big(\sup_{t\in[0,1]} |Z^{(n)}(t)| \ge C \big) &\le \mathbb{P}\Big(\sup_{0\le j\le n} \Big| \frac{K_j^{(n)} - j(j-1)/n}{\sqrt{n}} \Big| \ge C - 1 \Big) \\ &= \mathbb{P}\Big(\sup_{1\le j\le n-1} \Big| \frac{Y_j^{(n)}(n-j)}{\sqrt{n}} \Big| \ge C - 1 \Big). \end{aligned} \tag{15}$$

这里要注意，我们无需在上确界中考虑 $j = 0$ 和 $j = n$，因为$K_j^{(n)} - j(j-1)/n = 0$ 无论哪种情况都成立。由于 $(Y_j^{(n)})_{1\le j\le n-1}$ 是一个鞅，我们可以使用 Doob 的 L^p-不等式 [14, 定理 11.2]。对于任意实数 $C > 0$ 以及任意满足 $1 \le k \le n-1$的固定整数 k，我们有

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

至此，我们已具备证明该界所需的全部前提。我们将 (15) 中取上确界的区间进行划分，应用鞅不等式，然后对所有这些上界求和，便可得到所需结果。对于

每个整数 $i>0$，令 $I_i^{(n)} := [2^{-i}n, 2^{-i+1}n] \cap \mathbb{Z}$。我们得到

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

其中，在最后一个不等式中，我们将方差估计如下：

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

最后，结合并集界以及观察到的事实 $\sum_{i\geq 1}2^i/(C-1)^2=4/(C-1)^2$，可得 (14) 中的上界，从而完成证明。□

接下来，我们考察有限维分布的行为。

引理 5. 设 r 为固定正整数，且 $\mathbf{t}=(t_1,\ldots,t_r)\in(0,1)^r$。那么，对于 $n\to\infty$，随机向量

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

经过中心化和重新缩放后，当 $n\to\infty$ 时，它依分布收敛于一个多元正态分布，

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

其中期望向量 $\mathbb{E}\mathbf{K}^{(n)}_{\lfloor \mathbf{t}n\rfloor}$ 满足

$$\mathbb{E}\mathbf{K}^{(n)}_{\lfloor \mathbf{t}n\rfloor}=n(t_1^2,t_2^2,\ldots,t_r^2)+\mathcal{O}(1), \tag{16}$$

且方差协方差矩阵 $\Sigma=(\sigma_{i,j})_{1\leq i,j\leq r}$ 的元素为

$$\sigma_{i,j}=\begin{cases}t_i^2(1-t_j) & \text{if } i\leq j,\\ t_j^2(1-t_i) & \text{if } i>j.\end{cases} \tag{17}$$

证明。根据引理 2 以及凯莱关于大小为 n 的标号树的著名计数公式，我们得到.....的概率生成函数

分别是 $1 < j_1 < j_2 < \cdots < j_r < n$ 步后边增量的数量，其表达式为

$$\begin{aligned} P_n(z_1, z_2, \ldots, z_r) &= \frac{E_n(z_1, z_2, \ldots, z_r)}{n^{n-2}} \\ &= \prod_{i=1}^{r} \Big(1 - \frac{j_r}{n} + \frac{j_i}{n} z_i + \sum_{h=i+1}^{r} \frac{j_h - j_{h-1}}{n} z_h\Big)^{j_i - j_{i-1}}, \end{aligned} \tag{18}$$

其中 $j_0 = 1$，为方便起见。鉴于这个概率生成函数可以很好地分解，我们可以使用一个一般性结果，例如多维拟幂定理（参见 [10]），来证明相应的随机向量

$$\Delta_{\mathbf{j}}^{(n)} = (K_{j_1}^{(n)}, K_{j_2}^{(n)} - K_{j_1}^{(n)}, \ldots, K_{j_r}^{(n)} - K_{j_{r-1}}^{(n)})$$

经过适当缩放后收敛到多元高斯极限分布。不过，这里有一个简单的概率论论证：注意到 $\Delta^{(n)}$ j 可以看作 r 个独立的多项分布随机向量之和的边缘分布：写作 $t_i = j_i/n$，并考虑 $M_j \sim \mathbf{Multi}(j_i - j_{i-1}, \mathbf{p}_i)$，其中

$$\mathbf{p}_i = (p_{i,0}, p_{i,1}, \ldots, p_{i,r}) \in [0,1]^r \quad \text{such that} \quad p_{i,h} = \begin{cases} 1 - t_r & \text{if } h = 0, \\ 0 & \text{if } 0 < h < i, \\ t_i & \text{if } h = i, \\ t_h - t_{h-1} & \text{otherwise.} \end{cases} \tag{19}$$

根据构造，M_i 的概率生成函数为

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

因此，和 $M_1 + \cdots + M_r$ 的概率生成函数是一个乘积，它与式（18）非常相似（事实上，当我们令 $z_0 = 1$时二者相等，这相当于对第一分量取边缘分布）。为了使下面的论证在形式上更易读，而且由于第一分量与我们完全无关，我们稍微滥用一下记号，用 M_i 表示 $1 \le i \le r$ 所对应的边缘化多项分布。

为方便起见，我们稍作调整：不再固定整数向量$\mathbf{j} = (j_1, \ldots, j_r)$，而是固定 $\mathbf{t} = (t_1, \ldots, t_r)$，其中 $0 < t_1 < \cdots < t_r < 1$，并定义$\mathbf{j} = \lfloor \mathbf{t} n \rfloor$。这里，$n$ 被视为足够大，从而对应的整数向量 $1 < \lfloor t_1 n \rfloor < \cdots < \lfloor t_r n \rfloor < n$仍满足这些条件。

根据多元中心极限定理，众所周知，多项分布随机向量 $M \sim \mathbf{Multi}(n, \mathbf{p})$ 在 $n \to \infty$ 及之后

经过适当缩放后，依分布收敛于多元正态分布，

$$\frac{M - n\mathbf{p}}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathrm{diag}(\mathbf{p}) - \mathbf{p}^\top \mathbf{p}). \tag{20}$$

因此，我们发现

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

其中方差-协方差矩阵由下式给出

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

通过对部分和进行简单的（线性）变换，增量随机向量 $\Delta^{(n)}_{\lfloor \mathbf{t}n\rfloor}$ 可转化为$\mathbf{K}^{(n)}_{\lfloor n\rfloor \mathbf{t}}$。这证明，经过中心化和重新缩放后，$\mathbf{K}^{(n)}_{\lfloor \mathbf{t}n\rfloor}$ 收敛于多元正态分布。

相应方差-协方差矩阵的元素，既可以通过考虑部分求和，从 $t_1\Sigma_1 + \cdots + (t_r - t_{r-1})\Sigma_r$ 的元素机械地确定；也可以利用我们对鞅 $Y_{(n)\,j}$ 的观察来确定。特别地，利用（12），对于固定的 $s, t \in [0,1]$ 且满足 $s < t$，我们得到

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

其中我们利用了鞅的性质，以及 $Y^{(n)}_j$ 的二阶矩等于方差 $n^{-2}j(j-1)/(n-j)$ 这一事实。最终，这验证了（17），从而完成了证明。□

要证明过程序列 $(Z^{(n)})_{n\geq 1}$ 是紧的，最后还需要对相应路径的增长给出一个界。

引理 6. 存在一个常数 λ，使得对于所有s、$t\in[0,1]$ 以及所有 $n\in\mathbb{Z}_{>0}$，以下不等式均成立：

$$\mathbb{E}\big(\big|Z^{(n)}(t)-Z^{(n)}(s)\big|^4\big)\le\lambda\,|t-s|^2. \tag{21}$$

证明。我们先考虑 $sn=\ell$ 和 $tn=m$ 均为整数的情形。假设 $\ell>m$。参照引理 5 的证明中的论证可知，随机变量 $K_\ell^{(n)}-K_m^{(n)}$的分布与两个独立多项分布之和中的最后一个分量相同，因此

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

我们将中心化二项分布记作 Bin*(n, p)，即从二项分布 Bin(n, p) 中减去其均值 np。于是有

$$K_\ell^{(n)}-K_m^{(n)}-\frac{\ell^2-m^2}{n}\sim\text{箱}^*\left(m-1,\frac{\ell-m}{n}\right)+\text{箱}^*\left(\ell-m,\frac{\ell}{n}\right)-\frac{\ell-m}{n}. \tag{22}$$

服从 Bin*(n, p) 的随机变量的四阶矩，也就是服从 Bin(n, p) 的随机变量的四阶中心矩，为

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

注意，对于式 (22) 中的 Bin* 随机变量，我们有 $\frac{(m-1)(\ell-m)}{n}\le\ell-m$ 且 $\frac{(\ell-m)}{n}\ell\le\ell-m$。因此，这两个中心化二项随机变量的四阶矩均有如下上界：

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

因此，$K_\ell^{(n)}-K_m^{(n)}-\frac{\ell^2-m^2}{n}$ 是三个随机变量之和（其中第三个实际上是常量），它们的四阶矩分别有如下上界：$4(\ell-m)^2$, $4(\ell-m)^2$ 和 $(\ell-m)^4n^{-4}\le(\ell-m)^2$。应用不等式 $\mathbb{E}((X+Y+Z)^4)\le 27(\mathbb{E}(X^4)+\mathbb{E}(Y^4)+\mathbb{E}(Z^4))$，该不等式是詹森不等式的一个简单推论，可得

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

因此，如果 $tn=\ell$ 和 $sn=m$ 是整数，我们有

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

其次，考虑 tn 和 sn 位于两个连续整数 $\tilde{m}$和 $m+1$之间的情况：$\tilde{s}n=m\le sn\le tn\le m+1=\tilde{t}n$。在这种情况下，我们可以将

差值 $Z^{(n)}(t) - Z^{(n)}(s)$（根据 $\tilde{K}_t^{(n)}$ 的定义使用线性插值）表示为

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

第一项的四阶矩为

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

因为 $|t-s| \le |\tilde{t}-\tilde{s}| = \frac{1}{n}$。同样，容易看出第二项的四次方不超过 $(t-s)^2$。因此，初等不等式 $\mathbb{E}((X+Y)^4) \le 8(\mathbb{E}(X^4) + \mathbb{E}(Y^4))$ 给出

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

最后，在一般情形下，设 t 和 s 是区间 $[0,1]$ 内的任意实数，且 tn 和 sn 不位于相邻整数之间，我们可以写成

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

对于某些实数 u_1, u_2，满足 $s \le u_2 \le u_1 \le t$，使得 $u_1 n$ 和 $u_2 n$ 为整数，并且 $tn \le u_1 n + 1$ 以及 $sn \ge u_2 n - 1$。结合上面的界，并再次使用不等式 $\mathbb{E}((X+Y+Z)^4) \le 27(\mathbb{E}(X^4)+\mathbb{E}(Y^4)+\mathbb{E}(Z^4))$，我们得到

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

使用 $\lambda = 52704$ 完成引理的证明。 □

现在剩下的就是结合这两个要素，证明关于极限过程的主要结果。

定理 3 的证明。 该证明依赖于一个众所周知的结果：如果相应概率测度序列具有紧性，并且有限维概率分布收敛，则随机过程序列将收敛到一个极限过程（参见 [5, 定理 7.1、定理 7.5]）。

根据引理 4 中的初始分布紧性和引理 6 中的矩界，可推出紧性（见 [14, 定理 13.29、21.42]）。引理 5 刻画了原过程 $(K_{\lfloor tn\rfloor}^{(n)})_{t\in[0,1]}$ 的有限维分布的（极限）行为。应用斯卢茨基定理 [14, 定理 13.18]，并注意到

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

机械计算可得 $\mathbb{E}((K^{(n)}_{\lfloor tn\rfloor+1} - K^{(n)}_{\lfloor tn\rfloor})^2) = \mathcal{O}(1)$（这也可由引理 6 得出）。

注意，由于有限维分布收敛于高斯分布，极限过程 $(Z^\infty(t))_{t\in[0,1]}$ 本身也是高斯过程——这意味着它完全由一阶和二阶矩刻画。由引理 5 可知，对于所有 $s, t \in [0,1]$ 且满足 $s < t$ 的情形，有

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

可以验证，若 $(W(t))_{t\in[0,1]}$ 是标准维纳过程，则高斯过程 $((1-t)W(t^2/(1-t)))_{t\in[0,1]}$ 具有相同的一阶和二阶矩，因此其分布也与 Z^∞ 相同。□

3. 根簇大小

现在，我们将注意力从未揭示边的数量转向揭示过程中图内出现的连通分量（或簇）的大小。将树模型改为有根标号树会很方便，因为有根树的性质使我们能够专注于某个特定的簇——包含根顶点的簇。如果根顶点尚未揭示，我们就将根簇的大小视为 0。形式上，令随机变量 $R^{(k)}_n$ 表示大小为 n 、含有 k 个未揭示顶点的（均匀）随机有根标号树的根簇大小。

利用标记结构的符号方法（参见 [8, 第 II 章]），我们可以为相应的组合类建立形式化描述，进而导出相关生成函数的函数方程。令 $\mathcal{T}^\bullet$ 表示有根标号树类，令 $\mathcal{G}$ 表示对 $\mathcal{T}^\bullet$ 的细分，其中顶点可以是已覆盖或未揭示的，未揭示顶点用标记 U 标记。最后，令 $\mathcal{F}$ 表示对 $\mathcal{G}$ 的进一步细分，其中根簇中的所有未揭示顶点还额外用标记 V 标记。采用直接的“自顶向下方法”，即根据根顶点对树族中的成员进行分解，可得形式化描述

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

请注意， $\mathcal{F}$ 形式化描述中的第一项对应于根顶点已覆盖的情况，因此根簇大小为零。

引入相应的指数生成函数 $F := F(z,\ u,\ v)$ 和 $G := G(z,\ u)$,

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

$$G(z,u) := \sum_{T\in\mathcal{G}} \frac{z^{|T|}u^{\#U \text{ in } T}}{|T|!} = \sum_{n\geq 1}\sum_{0\leq k\leq n} \frac{n^{n-1}}{n!}\binom{n}{k} z^n u^k,$$

我们得到刻画方程

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

当然，$G(z,u) = T^\bullet(z(1+u))$，其中 $T^\bullet$ 是与 $\mathcal{T}^\bullet$——凯莱树函数——相关联的指数生成函数。从 (23) 出发，可以推导出关于 $R_n^{(k)}$ 的以下结果。

定理7。期望 $\mathbb{E}(R_n^{(k)})$ 对于 $0 \le k \le n$ 和 $n \ge 1$，由下式给出：

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

根据 $k = k(n)$ 的增长情况，$\mathbb{E}(R_n^{(k)})$ 具有以下渐近行为：

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

证明。引入 $E := E(z,u) = \frac{\partial}{\partial v} F(z,u,v)\big|_{v=1} = \sum_{n,k} \frac{n^{n-1}}{n!} \binom{n}{k} \mathbb{E}(R_n^{(k)})$ 后，对 (23) 关于 v 求偏导数即可得到

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

利用拉格朗日反演公式提取 E 的系数（例如参见 [8, 定理 A.2]），可得

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

并且进一步

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

利用 $\mathbb{E}(R_n^{(k)}) = \frac{[z^n u^k]E}{[z^n u^k]G} = \frac{[z^n u^k]E}{n!\, n^{n-1}} \binom{n}{k}$，我们得到式 (24) 中所述的结果：

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

为了分析 $\mathbb{E}(R_n^{(k)})$ 的渐近行为，以下积分表示很有用，

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

利用伽马函数的积分表示，可以很容易地验证：

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

为了渐近地计算积分 (25)，我们首先证明，对于范围 $x \geq n^{\frac{1}{2}+\epsilon}$，任意固定的 $\epsilon > 0$，该部分对积分的贡献呈指数级小，因此可以忽略。具体而言，考虑被积函数并令 $t = \frac{x}{n}$，可得以下估计，该估计对 $k \in [0, n]$ 一致成立：

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

通过简单的单调性考察可得，对于 $t \geq 1$（因此 $x \geq n$），有 $t - \ln(1+t) \geq \frac{t}{4}$；对于 $t \in [0, 1]$（因此 $x \in [0, n]$），有 $t - \ln(1+t) \geq \frac{t^2}{4}$。利用这些估计并计算所得积分，我们得到相应范围内该积分的以下界：

$$\begin{aligned}\int_n^\infty (x-1) e^{-x} \Big(1 + \frac{x}{n}\Big)^k dx &\leq \int_n^\infty x e^{-\frac{x}{4}} dx = 16\big(1 + \frac{n}{4}\big) e^{-\frac{n}{4}}, \\ \int_{n^{\frac{1}{2}+\epsilon}}^n (x-1) e^{-x} \Big(1 + \frac{x}{n}\Big)^k dx &\leq \int_{n^{\frac{1}{2}+\epsilon}}^\infty x e^{-\frac{x^2}{4n}} dx = 2n e^{-\frac{1}{4} n^{2\epsilon}},\end{aligned}$$

因此，结合这两种情况，对于 $k \in [0,n]$ 和 $\epsilon \in (0,\frac{1}{2}]$，结论一致成立：

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

此外，我们注意到（粗略地说），如果 k 与 n 相距足够远，则可以扩展积分区间，而不会引入显著贡献。具体而言，令 $\delta = 1-\frac{k}{n}$ 和 $x = nt$，我们得到被积函数

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

其中，对于 $t \ge 0$，我们使用了平凡估计 $\ln(1+t) \le t$。例如，如果 $\delta \ge n^{-\frac{1}{4}}$，即 $k \le n - n^{\frac{3}{4}}$，则可由该估计轻易得到，区间 $x \ge n^{\frac{1}{4}+\epsilon}$ 对积分的贡献是渐近可忽略的：

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

为了得到定理中所述积分的渐近表达式，我们将考虑 k 相对于 n 的增长，对区间 $x \le n^{\frac{1}{2}+\epsilon}$（或分别对 $x \le n^{\frac{1}{4}+\epsilon}$）上的被积函数作适当展开，并计算所得积分；其中我们应用“尾部交换技巧”，即可以将积分区间扩展至 $x \ge 0$，因为这样只会增加渐近可忽略的贡献。

- k较小或位于中心区域：假设 $k \le n - n^{\frac{3}{4}}$，将被积函数在 $x \le n^{\frac{1}{4}+\epsilon}$处展开，可得如下展开式（在此范围内有一致界）：

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

从而得到积分的如下计算结果：

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

这里我们使用了公式

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

当然，这尤其给出

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

• k 亚临界大：下面我们讨论 $\frac{k}{n} \to 1$ 的情形；在这种情况下，必须根据差值$d = n - k$的增长行为加以区分。我们先考察区域 $\sqrt{n} \ll d \ll n$，即 $d = o(n)$, 但$d = \omega(\sqrt{n})$，对此我们得到被积函数的如下展开式（对于 $x \le n^{\frac{1}{2}+\epsilon}$）：

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

考虑相应的积分（并应用尾部交换），我们得到

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

利用 (26) 以及

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

我们得到所述结果：

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

• k 临界大：当差值 $d = n - k$ 的阶为$\Theta(\sqrt{n})$时，我们得到被积函数的如下展开式：

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

因此，完成所涉及的积分后，我们得到

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

由于，对于 $\ell \ge 0$,

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

这就得出

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

令 $d = c\sqrt{n}$，并应用代换 $t = c + \frac{x}{\sqrt{n}}$，计算该积分，得到所述结果:

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

- k 超临界大：对于 $d = n - k = o(\sqrt{n})$，我们得到展开式

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

使用与前面类似的计算,

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

得到所述结果:

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

□

我们甚至可以得到 $R_n^{(k)}$ 的精确分布。我们想简要概述两种不同的方法：首先，生成函数 $F = F(z,u,v)$ 的显式公式既可以通过处理递归描述式 (23) 得到，也可以直接将 $\mathcal{F}$ 分解为一棵构成未覆盖根簇的树，并附上一个带有被覆盖的根的森林。无论采用哪种方法，都可得到

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

注意，第二个加数 $\frac{G}{1+u} = ze^G$ 对应于根顶点被覆盖的情形。应用拉格朗日反演公式提取系数后，可得 $F_{n,\,k,\,m} := n![z^n u^k v^m]F(z,\ u,\ v)$ 的显式公式，即具有 n 个顶点、其中 k 个未被覆盖且 m 个属于根簇的标号有根树的数量（当 $0 \leq m \leq k \leq n$ 和 $n \geq 1$ 时）：

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

根据该公式，可以直接得到定理 9 中给出的概率 $\mathbb{P}(R_n^{(k)} = m) = \frac{F_{n,\,k,\,n^{n-1}}}{\binom{n}{k}}^m$。这些计算虽然直接，但略显冗长,

因此我们省略这些计算，因为下文将以一种更一般、更优雅的方式推导出这些结果。

或者，也可以采用一种更具组合性的方法来确定这些概率：有一个初等公式可以计数满足指定顶点集构成一个簇的树。

引理 8. 设 n 和 k 为正整数，且 $r_1, r_2, \ldots, r_\ell$ 为满足 $r_1 + \cdots + r_\ell \le k$ 的固定正整数。此外，取 $[k]$ 的两两不相交子集 $R_1, \ldots, R_\ell$，使得对所有 i 都有 $|R_i| = r_i$。满足以下条件的 n 顶点标号树的数量为：$R_1, \ldots, R_\ell$ 是由标签属于 $[k]$ 的顶点所诱导的森林的连通分量。

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

证明。我们将每棵这样的树 T 视为具有 n 个顶点的完全图 K 的一棵生成树。令 $r = r_1 + \cdots + r_\ell$。$K$ 的顶点可划分为集合 $R_1, \ldots, R_\ell$、在 $\{1, 2, \ldots, k\}$ 中剩余的 $k - r$ 个顶点组成的集合 Q，以及标签大于 k 的 $n - k$ 个顶点组成的集合 S。首先注意，由集合 $R_1, \ldots, R_\ell$ 所诱导的连通分量有 $r_1^{r_1-2}\cdots r_\ell^{r_\ell-2}$ 种选取方式。若将对应于 $R_1, \ldots, R_\ell$ 的顶点集收缩为单个顶点 $v_1, \ldots, v_\ell$，则 K 成为一个具有 $n - r + \ell$ 个顶点的多重图 K'，而树 T 收缩后成为 K' 的一棵生成树 T'。注意，从 v_i 到 K' 中每个其他顶点都有 r_i 条边，并且 T' 不能包含从 v_i 到 $\{v_1, \ldots, v_\ell\} \cup Q$ 中另一个顶点的任何边。因此，从 K' 中移除所有这类边，得到多重图 K'' 后，T' 仍是一棵生成树。反过来，若任取 K'' 的一棵生成树，并分别用 $R_1, \ldots, R_\ell$ 的生成树替换顶点 $v_1, \ldots, v_\ell$，就得到一棵具有 n 个顶点且满足所需性质的标号树。接下来需要计数 K'' 的生成树；它的邻接矩阵具有如下分块形式

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

此处，$\mathbf{1}$ 表示全为 1 的（列）向量，$\mathbf{r}$ 表示其元素为 r_1、……、r_ℓ 的（列）向量，O 表示全为 0 的矩阵，E 表示全为 1 的矩阵，I 表示单位矩阵。这些分块分别对应于 ℓ、$k - r$ 和 $n - k$ 行/列。现在可以通过矩阵树定理确定生成树的数量：拉普拉斯矩阵为

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

其中 D 是一个对角矩阵，其对角元素为 r_1，…，r_ℓ。我们的目标是计算 L 删除第一行和第一列后的行列式；将该矩阵记为 L_1。如果从最后几行中每一行减去 $\frac{1}{n-k}$ 倍的前 $\ell - 1$ 行，

$n-k$ 行 L_1，我们得到一个矩阵，其中前 $\ell-1$ 列里除对角元素外的所有元素均为 0。因此，行列式等于这些对角元素 $r_2(n-k)$ ...， $r_\ell(n-k)$的乘积，再乘以一个分块形式的矩阵的行列式

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

其中各块的长度分别为 $k-r$ 和 $n-k$。该矩阵以 $n-r$ 为特征值，其重数为 $k-r-1$，因为减去 $n-r$ 倍的单位矩阵后，会得到一个有 $k-r$ 行相同的矩阵。出于同样的原因， n 是重数为 $n-k-1$ 的特征值。接下来需要确定其余两个特征值。相应的特征向量可按如下方式构造：令前 $k-r$ 个分量（对应第一块）都等于 a，其余分量都等于 b。容易验证，若满足联立方程，则该向量是特征值为 λ 的特征向量

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

。该方程组的 2×2 阶系数矩阵的两个解就是这两个特征值，其乘积等于该 2×2 矩阵的行列式，即

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

最终可得， L_1 的行列式，即 K'' 的生成树数量，等于

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

乘以 $r_1^{r_1-2}\cdots r_\ell^{\ell-2}$ （由连通分量 R_1、.....、 R_ℓ诱导的生成树的可能数量），即可得到所需公式。□

由针对 $\ell=1$ 的引理 8 可知，概率 $\mathbb{P}(R_n^{(k)}=r)$ 可通过将 $n^{n-k-1}(n-r)^{k-r-1}(n-k)r^{r-1}$ 与 $r\binom{k}{r}$ 相乘来得到（这给出了根包含在前 k 个未揭示顶点中大小为 r 的簇内的 n个顶点上的有根标号树的数量），然后再除以 n^{n-1}，即 n 个顶点上的标号有根树的数量，进行归一化。

定理 9。 $R_n^{(k)}$的精确分布由以下概率质量函数（概率质量函数）刻画，该函数对 $0\le m\le k\le n$ 和 $n\ge 1$ 由下式给出（其他情况下等于0 ）：

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

根据 $k=k(n)$ 的增长情况，我们得到以下极限行为 r:

- k 较小，即，$k=o(n)$:
$$R_n^{(k)} \xrightarrow{p} 0.$$
- k 在中心区域中，即， $k\sim\alpha n$ 具有 $0<\alpha<1$:

$R_n^{(k)} \xrightarrow{d} R_\alpha$，其中离散随机变量 R_α 由其概率质量函数刻画。
$$\mathbb{P}(R_\alpha=m)=:p_m=\begin{cases}1-\alpha, & m=0,\\ \frac{m^m}{m!}(1-\alpha)\alpha^m e^{-\alpha m}, & m\geq 1,\end{cases}$$
或者也可以用概率生成函数 $p(v)=\sum_{m\geq 0}p_m v^m=1-\alpha 1-T^\bullet(v\alpha e^{-\alpha})$ 来刻画。

- k 亚临界大，即 $k=n-d$，其中 $d=\omega(\sqrt{n})$ 且 $d=o(n)$:
$$\left(\frac{d}{n}\right)^2\cdot R_n^{(k)} \xrightarrow{d} \textsc{Gamma}\left(\frac{1}{2},\frac{1}{2}\right),$$
其中，伽马$(1\,2,1\,2)$是一个伽马分布，其密度函数为$f(x)=\frac{1}{\sqrt{2\pi x}}e^{-\frac{x}{2}}$，其中 $x>0$。
- k 临界大，即 $k=n-d$，其中 $d\sim c\sqrt{n}$ 且 $c>0$:
$$\frac{1}{n}\cdot R_n^{(k)} \xrightarrow{d} R(c),$$
其中，连续随机变量 $R(c)$ 由其密度函数 $f_c(x)=\frac{1}{\sqrt{2\pi}\sqrt{x}}\frac{c}{(1-x)^{\frac{3}{2}}}e^{-\frac{c^2x}{2(1-x)}}$ 表征，其中 $0<x<1$。
- k 超临界大，即 $k=n-d$，其中 $d=\omega(1)$ 且 $d=o(\sqrt{n})$:
$$\frac{1}{d^2}\cdot\left(n-R_n^{(k)}\right) \xrightarrow{d} D,$$
其中，连续随机变量 D 由其密度函数 $f(x)=\frac{1}{\sqrt{2\pi}\,x^{\frac{3}{2}}}e^{-\frac{1}{2x}}$ 表征，$x>0$。
- k 超临界大且具有固定差值，即 $k=n-d$ 且 d 固定:
$$n-d-R_n^{(k)} \xrightarrow{d} D(d),$$
其中离散随机变量 $D(d)$ 由其概率质量函数刻画。
$$\mathbb{P}(D(d)=j)=:p_j=e^{-d}\cdot\frac{d(d+j)^{j-1}}{j!}\cdot e^{-j},\quad j\geq 0,$$
或者也可以通过概率生成函数 $p(v)=\sum_{j\geq 0}p_j v^j=e^{d(T^\bullet(\frac{v}{e})-1)}$ 来表示。

证明。 $R_n^{(k)}$的概率质量函数可由定理陈述之前的相关论述得到。由于其具有显式形式，在区分若干种情形后应用阶乘的斯特林公式，便可相当直接地得到定理 9 中所述的极限分布

通过区分几种情况并对阶乘应用斯特林公式的方式。□

注记 4. 当然，对于标号树， $R_n^{(k)}$的分布与随机顶点的簇大小分布相同。此外，通过条件化，可以轻松地将定理 9 的结果转化为关于第 k 个未揭露顶点的簇大小 $S_n^{(k)}$的结果： $\mathbb{P}(S_n^{(k)} = m) = \mathbb{P}(R_n^{(k)} = m | R_n^{(k)} > 0) = \frac{n}{k} \cdot \mathbb{P}(R_n^{(k)} = m)$，对于 $m \geq 1$。

4. 最大未揭露连通分量的大小

掌握了根簇的行为后，我们回到无根标号树，并研究最大簇的大小。为此，我们引入随机变量 $X_{n,r}^{(k)}$，它表示在一棵大小为 n 的均匀随机标号树中，揭示顶点 1 至 k 后，大小为 r 的连通分量数量。

形式上， $X_{n,r}^{(k)}: \mathcal{T}_n \to \mathbb{Z}_{\geq 0}$。注意，对于所有标号树 $T \in \mathcal{T}_n$，我们都有

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

定理 10. 设 $n, k, r \in \mathbb{Z}_{\geq 0}$ 满足 $0 \leq r \leq k \leq n$。从一棵大小为n 、均匀随机选取的标号树中揭示 k 个顶点后，大小为 r 的连通分量的期望数量为

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

*P*证明。 注意到 $X_{n,r}^{(k)}$ 可以写成伯努利随机变量之和 s

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

其中 $X_{n,S}^{(k)}$ 取值为 0 或 1，具体取值取决于在经过 k 步揭示后， S 中的顶点是否构成一个簇。根据对称性和期望的线性性质，我们有

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

右侧期望值的公式可由引理 8 得出，从而证明该定理。□

秉承式 (28) 中观察的思路，引理 8 中的公式为以下求和恒等式提供了组合证明。

推论 11. 设 $n, k \in \mathbb{Z}_{\geq 0}$ 满足 $0 \leq k \leq n$。则恒等式

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

成立。

证明。 右边统计了所有含有 n 个顶点的标号树中，$[k]$ 中的顶点总数。左边进行相同的统计，其中各项求和统计的是大小为 r 的连通分量中的顶点数。□

注 5。 注意，恒等式 (30) 可以改写为

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

这是阿贝尔二项式定理的一种特例形式——这是一个经典且广为人知的结果；例如参见 [24]。

对于一棵树 $T \in \mathcal{T}$，令 $c^{(k)}_{\max}(T)$ 表示揭示前 k 个顶点后，T 中最大的连通分量。

定理 12. 设 $n \in \mathbb{Z}_{\geq 0}$，并设 $T_n \in \mathcal{T}_n$ 是均匀随机选取的一棵树。那么，随机变量 $c^{(k)}_{\max}(T_n)$ 在 $n \to \infty$ 时的行为可描述如下：

- 当 $k = n - d$ 且 $d = \omega(\sqrt{n})$（次临界情形）时，我们有 $c_{(k)\max}(T_n)/n \xrightarrow{p} 0$。
- 当 $k = n - d$ 且 $d = o(\sqrt{n})$（超临界情形）时，我们有 $c_{(k)\max}(T_n)/n \xrightarrow{p} 1$。以高概率，存在一个“巨型”连通分量，其大小渐近等于 n。

证明。 对于次临界情形，我们使用定理 7 中的期望根簇大小。由于大小为 r 的簇包含根的概率为 $\frac{r}{n}$，因此我们有

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

这意味着

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

因此，若对于任意固定的 $\epsilon > 0$，都有 $m = \epsilon n$，则有 $\mathbb{P}(c^{(n-d)}_{\max}(T_n) \geq m) \to 0$。

在超临界情形下，我们回顾定理7中关于根簇大小的对应情形。利用马尔可夫不等式可得，对任意 $\epsilon > 0$,

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

因此，以高概率，根簇是大小为 $\sim n$ 的最大簇。将这一结果从有根树推广到无根树，即可证明该定理。□

在临界情形下，当 $n-k \sim c\sqrt{n}$ 对某个常数 c 成立时，我们还能够刻画 $c_{\max}^{(k)}(T_n)/n$ 的行为：该变量依分布收敛于一个连续极限分布。为了描述这一分布，我们先需要以下引理。

引理 13. 假设 $n-k \sim c\sqrt{n}$，并固定 $\overline{\alpha} > 0$。当 $n \to \infty$ 时，均匀随机标号树 T_n（含有 n 个顶点）中，由标签属于 $[k]$ 的顶点诱导的森林包含两个连通分量（每个至少有 αn 个顶点且大小相等）的概率趋于0 。

证明。假设存在两个连通分量，每个含有 a 个顶点，其中 $a \geq \alpha n$。根据引理 8，发生这种情况的概率为

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

前提是 $k \geq 2a$（否则，该概率显然为 0）。开头的多项式系数给出了选择两个连通分量标签的方式数，分母则是标号树的总数。我们的目标是估计这个表达式。$k = 2a$ 的情形较易单独处理，因此假设 $k > 2a$。于是，根据斯特林公式，对于某个常数 C_1，我们有

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

众所周知，当 β 固定时，$(1+\beta/x)^x$ 随 x 增大而增大。因此，如果 $k-2a \leq n^{3/4}$，根据假设 $n-k \sim c\sqrt{n}$，我们有：

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

在这种情况下，$P(a)$ 比 n 的任何幂次趋于 0 的速度都快。否则，即如果 $k-2a > n^{3/4}$，则有 $k-2a \sim n-2a$，并利用 $a \geq \alpha n$ 以及与上述

相同的泰勒展开，我们得到

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

其中 C_2 为常数。我们可以将其改写为

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

具有 $f(x) = x^{3/2}e^{-x/2}$。由于此函数有界，我们已证明 $P(a) = \mathcal{O}(n^{-2})$，且对 a一致。对 a 的所有可能取值求和可知，所讨论的概率为 $\mathcal{O}(n^{-1})$。特别地，它趋于 0。□

现在我们可以证明以下关于 $c_{\max}^{(k)}(T_n)/n$ 的极限分布的描述。

定理14。 假设 $n-k \sim c\sqrt{n}$，并固定 $\alpha > 0$。在一棵具有 n 个顶点的均匀随机标号树T_n 中，由标签属于 $[k]$ 的顶点诱导的森林包含一个至少有 αn 个顶点的连通分量的概率趋于

$$\sum_{j\ge 1} \frac{(-1)^{j-1}c^j}{(2\pi)^{j/2}} \int\limits_{\substack{\alpha \le t_1 < \cdots < t_j \\ \tau_j = t_1+\cdots+t_j < 1}} \cdots \int \prod_{i=1}^{j} t_i^{-3/2}(1-\tau_j)^{-3/2} \exp\Big(-\frac{c^2\tau_j}{2(1-\tau_j)}\Big)\, dt_1 \cdots dt_j$$

当 $n \to \infty$ 时。

证明。 设 $r_1, \dots, r_\ell$为正整数，且满足 $\alpha n \le r_1 < \cdots < r_\ell$和 $r_1+\cdots+r_\ell \le k$。根据引理 8，标签属于 $[k]$ 的顶点所诱导的森林，其连通分量大小为 $r_1, \dots, r_\ell$的概率由下式给出，其中$r = r_1 + \cdots + r_\ell$:

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

并且，采用与引理 13 中相同的论证可知，该概率为 $\mathcal{O}(n^{-\ell})$，且在 r_1、……、r_ℓ 中一致成立。此外，若令 $r_i = t_i n$，根据斯特林公式，经过一些整理后可得该概率的如下渐近公式：令$t = t_1 + \cdots + t_\ell$，则为

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

此外，根据容斥原理，至少存在一个大小不小于 αn 的连通分量的概率为

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

最后一个误差项考虑了存在两个大小相同的分支的可能性。根据引理 13，这一事件的概率为 $\mathcal{O}(n^{-1})$。请注意，我们实际上只需要有限项，因为当 $j\alpha > 1$ 时，求和为空。将 $P(r_1,\ldots,r_\ell)$ 的渐近公式代入并取极限后，求和便变为积分，从而得到所需的公式。□

致谢

我们感谢 Svante Janson 指出本文扩展摘要中定理 3 的证明存在一个漏洞。

参考文献

[1] L. Addario-Berry、N. Broutin 和 C. Holmgren。用链锯马尔可夫链砍伐树木。《应用概率年刊》，24(6):2297–2339，2014。[2] D. Aldous 和 J. Pitman。标准加性合并过程。《概率年刊》，26(4):1703–1726，1998。[3] J. Bertoin。随机碎裂与凝聚过程，第102卷，收录于剑桥高等数学研究丛书。剑桥大学出版社，剑桥，2006。[4] G. Berzunza Ojeda 和 C. Holmgren。条件稳定伽尔顿–沃森树所导出的碎裂过程的不变性原理。预印本，arXiv:2010.07880。[5] P. Billingsley。概率测度的收敛。约翰·威立父子公司，纽约，第二版，1999。[6] P. Chassaing 和 G. Louchard。停车区块、布朗 excursion 与合并的相变。《随机结构与算法》，21(1):76–119，2002。[7] J. A. Fill、N. Kapur 和 A. Panholzer。极简单树的破坏。《算法学》，46(3-4):345–366，2006。[8] P. Flajolet 和 R. Sedgewick。解析组合学。剑桥大学出版社，剑桥，2009。[9] B. Hackl、A. Panholzer 和 S. Wagner。揭露一棵随机树。载于 M. D. Ward 编，第*33*届概率、组合与渐近算法分析国际会议（*AofA 2022*），第225卷，收录于莱布尼茨国际信息学会议录（*LIPIcs*），第10:1–10:17页，德国达格施图尔，2022。达格施图尔宫–莱布尼茨信息学中心。[10] C. Heuberger 和 S. Kropf。高维拟幂定理与贝里–埃森不等式。《数学月刊》，187(2):293–314，2018。[11] S. Janson。确定性树与随机树中的随机切割和记录值。《随机结构与算法》，29(2):139–179，2006。

[12] S. Janson。利用完整子区间排序以及随机演化序列中的最大游程数。《组合年刊》，12(4):417–447，2009。[13] J. F. C. Kingman。合并过程。《随机过程应用》，13(3):235–248，1982。[14] A. Klenke。概率论——综合教程。Universitext。Springer，Cham，2020。第三版。[15] M. Kuba 和 A. Panholzer。通过随机切割在有根树中隔离叶节点。《组合年刊》，12(1):81–99，2008。[16] J. L. Martin 和 V. Reiner。若干加权生成树计数函数的因式分解。《组合理论杂志*A*辑》，104(2):287–300，2003。[17] A. Meir 和 J. W. Moon。砍伐随机树。《澳大利亚数学学会杂志》，11:313–324，1970。[18] G. Miermont。由稳定树导出的自相似碎裂。I. 按高度分裂。《概率论及相关领域》，127(3):423–454，2003。[19] G. Miermont。由稳定树导出的自相似碎裂。II. 按节点分裂。《概率论及相关领域》，131(3):341–375，2005。[20] J. Pitman。合并随机森林。《组合理论杂志*A*辑》，85(2):165–193，1999。[21] J. Pitman。具有多重碰撞的合并过程。《概率年刊》，27(4):1870–1902，1999。[22] J. B. Remmel 和 S. G. Williamson。生成树与函数类。《电子组合学杂志》，9(1):研究论文34，24，2002。[23]D. Revuz 和 M. Yor。连续鞅与布朗运动，第293卷，收录于数学科学基础丛书。施普林格出版社，柏林，第三版，1999。[24] J. Riordan。组合恒等式。约翰·威立父子公司，纽约，1968。[25] P. Thévenin。稳定树上碎裂过程的几何表示。《概率年刊》，49(5):2416–2476，2021。